AD_____

Award Number: DAMD17-96-1-6254

TITLE: Computer-Aided Diagnosis and Feature-Guided Data Reduction Systems in Mammography

PRINCIPAL INVESTIGATOR: Heang-Ping Chan, Ph.D.

CONTRACTING ORGANIZATION: University of Michigan
Ann Arbor, Michigan 48103-1274

REPORT DATE: October 1999

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for public release;
Distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

20000907 137

DTIC QUALITY INSPECTED 4

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 1999	Annual (23 Sep 98 - 22 Sep 99)

4. TITLE AND SUBTITLE
Computer-Aided Diagnosis and Feature-Guided Data Reduction Systems in Mammography

5. FUNDING NUMBERS
DAMD17-96-1-6254

6. AUTHOR(S)
Heang-Ping Chan, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Michigan
Ann Arbor, Michigan 48103-1274

E-MAIL:
chanhp@umich.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

We have performed extensive evaluation of the computer detection programs and the GUI this year. The mass detection program has been evaluated with over 300 mammograms at the University of Michigan and the Georgetown University. The microcalcification detection program was evaluated with 260 mammograms. In a small-scale reading experiment simulating the pilot CAD reading of screening mammograms by four experienced mammography radiologists, we found that the CAD could improve the detection of cancer cases, but there might be a very small increase in the call-back rate. We expect that the pilot clinical study will provide information if the increase is statistically significant.

Two CADView workstations have been implemented at the University of Michigan and the Georgetown University. The pilot clinical study in our off-line screening mammography clinics has begun and will collect data for the analysis of the effects of CAD on radiologists' reading.

The CAD-guided image compression project is progressing as planned. The compression technique has been evaluated in a small data set described in the GU report last year. A large data set has been assembled and the preparation for the observer evaluation study has been completed. The subjective image quality comparison study is planned to start early next year.

Because of the change in the strategy for the CAD workstation development and the addition of the mass detection program, as described in the previous reports, as well as the incompatibility of different workstations and operating systems, there is a delay in starting the pilot clinical study. We have requested and obtained approval for a no-cost-time-extension of one year to make up for part of the work.

14. SUBJECT TERMS
Computer-aided diagnosis,breast cancer detection, data compression
Mammography

15. NUMBER OF PAGES
90

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

___ Where copyrighted material is quoted, permission has been obtained to use such material.

___ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

_X_ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

N/A In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

X For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

N/A In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

N/A In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

N/A In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

Chan Heng Pang 11/1/99

PI - Signature Date

(4) Table of Contents

(5) Introduction

In the United States, breast cancer is the leading cause of death in women between 40 to 55 years of age(1990). It is estimated that one out of eight women will develop breast cancer in their lifetime (Boring, et al. 1994, Harris, et al. 1992). There is considerable evidence that early diagnosis and treatment significantly improves the chance of survival for patients with breast cancer (Byrne, et al. 1994, Curpen, et al. 1995, Feig and Hendrick 1993, Moskowitz 1987, Seidman, et al. 1987, Smart, et al. 1995). The American Cancer Society — National Cancer Institute Breast Cancer Detection Demonstration Project (BCDDP) has shown that mammography contributes significantly in the detection of localized breast cancer in asymptomatic women (Seidman, et al. 1987).

Although mammography has a high sensitivity for detection of breast cancers when compared to other diagnostic modalities, studies indicate that radiologists do not detect all carcinomas that are visible on retrospective analyses of the images (Baines, et al. 1986, Bassett, et al. 1987, Bird, et al. 1992, Harvey, et al. 1993, Haug, et al. 1987, Hillman, et al. 1987, Kalisher 1979, Martin, et al. 1979, Moskowitz 1987, Wallis, et al. 1991). While double reading can reduce the miss rate in radiographic reading (Metz and Shen 1992, Thurfjell, et al. 1994), it also increases the cost of screening. In our ROC study (Chan, et al. 1990), we found that a CAD scheme, which alerts the radiologist to suspicious clusters of microcalcifications, can significantly improve radiologists' accuracy in detecting the microcalcifications under experimental conditions that simulate the rapid interpretation of screening mammograms. More recently, Kegelmeyer et al. (Kegelmeyer, et al. 1994) also showed that CAD can improve radiologists' detection of spiculated masses. These studies indicate that CAD is a viable alternative to double reading by radiologists.

Early breast cancers are often characterized by subtle clustered microcalcifications and masses (Tabar and Dean 1985). It has been reported that between 30 and 50% of breast carcinomas detected radiographically demonstrate microcalcifications on mammograms, and 40 to 50% of breast carcinomas present as masses. The high correlation between the presence of microcalcifications and masses and the presence of breast cancers indicates that an increase in the accuracy of detection and analysis of the characteristic features of these lesions may lead to further improvement in the efficacy of mammography as a screening procedure for the detection of early breast cancer.

In the past few years, we have been developing CAD algorithms in detection and classification of microcalcifications and masses using advanced image processing and computer vision techniques. Our CAD algorithms have provided very promising results in laboratory tests. At this stage, it is necessary to test the algorithms in a clinical trial with a large number of mammograms obtained from the general patient population before specific methods can be developed to further improve their performance. Therefore, our goals in this proposal are to implement our CAD algorithms in a fast workstation, develop user interfaces for efficient operation of the CAD programs, and conduct a pilot clinical trial of the CAD schemes at three mammographic screening sites. Based on the results of the pilot clinical trial, we can evaluate the sensitivity and specificity of the CAD algorithms, analyze the effects of the CAD schemes on mammographic screening, identify any potential problems in a clinical environment, and develop methods to further improve the CAD schemes in the future. We believe that this is a crucial step to develop a clinically practical CAD workstation.

It has been recognized that digital mammography is one of the key research areas for improvement in the diagnosis of breast cancer (Shtern, et al. 1995). Two of the major issues in digital mammography are the technological requirements in developing high resolution digital detectors and the transmission and archiving the large amount of data. A number of solid-state large-area digital detectors

are being developed for mammographic application. It has been generally recognized that a pixel size of no greater than 0.05 mm × 0.05 mm will be required for imaging the subtle features of microcalcifications. At this resolution, a single 8" × 10" mammogram will result in 40 MB of digital data.

Data compression can reduce the amount of data for transmission and storage. However, there is often a tradeoff between compression ratio and image fidelity. Data compression in mammography is especially difficult because of the very subtle image details such as microcalcifications and mass margins that need to be preserved. We have investigated the effects of data compression on computerized detection of microcalcifications previously. In the current proposal, we plan to develop a CAD guided data compression technique to maximize the compression efficiency with a minimum loss of information. Our approach is to preserve the original image information by lossless compression in potentially important regions on the mammograms indicated by the CAD programs. For breast areas outside these regions, we will apply the most efficient lossy compression technique that does not cause noticeable degradation of image details. We will conduct both receiver operating characteristic studies and subjective image quality ranking studies to compare observer performance on the uncompressed images, on images compressed with the selected lossy technique, and on images compressed with the standard JPEG technique.

The importance of this research is based on the fact that x-ray mammography is, at present, the most reliable diagnostic procedure for detection of early breast cancer. Our proposed research aims at the development of a CAD workstation which may assist radiologists in screening and characterizing abnormalities on mammograms and the development of an efficient CAD-guided data compression technique for digital mammography. The CAD workstation, once developed, can be implemented and operated cost-effectively in various breast imaging facilities as a second opinion, and thus will potentially increase the diagnostic accuracy of mammography for breast cancer detection. The data compression technique will facilitate the implementation of telemammography and digital mammography for breast cancer screening. These new technologies therefore are expected to have a significant impact on patient care, especially in rural and remote areas.

With the support of this grant from the USAMRMC Breast Cancer Research Program, we have developed a CAD workstation with a proper graphical user interface for a pilot clinical trial. CAD workstations have been implemented at the University of Michigan and at the Georgetown University. We have evaluated our mass and microcalcification detection programs with a large number of randomly sampled clinical cases. We have prepared cases for a subjective image quality comparison experiment to evaluate the feature guide data compression technique. Statistical methods are being developed for analysis of the pilot clinical data. We will discuss the details of these progresses in the following section.

(6) Body

During the funding period of 9/22/98 to 9/21/99, the three collaborating institutions in this Demonstration Project: University of Michigan, Georgetown University, and University of Iowa, have conducted the following tasks. The report from each of the institution is presented separately. A summary that links the tasks together and discusses the overall progress of the project is presented after the individual reports.

University of Michigan

(a) Evaluation of performance of computerized detection program

Preprocessing of mammograms

Before the input of a mammogram into the automated detection programs, the mammogram has to be processed to remove all the unexposed areas around the image, including the patient identification label. The breast image has also to be segmented from the non-breast background so that detection will be performed only within the breast area. We have been developing fully automated programs to perform these tasks. These programs are implemented in an UNIX Alphastation and is an integral part of the mass and microcalcification detection programs. We have tested the program in over 1000 unknown test mammograms to evaluate its accuracy in trimming and breast segmentation. Based on the evaluation of the unknown cases, we have modified the programs and the accuracy has been improved substantially. The programs currently could reliably trim off the unexposed film edges and segment the breast area in over 95% of the test cases that we collected at the University of Michigan. Because of the differences in the patient labels and screen-film cassettes, the programs had some problems in about 10% of the images collected at the Georgetown University. The current performance will be adequate for the pilot clinical study. However, further modifications are underway to improve the accuracy of segmentation for the mammograms.

Detection of Masses

The block diagram for the proposed detection scheme is shown in Fig. 1. When a digitized mammogram is input to the CAD system, edge trimming and breast area segmentation is applied to the image. Global density-weighted contrast enhancement (DWCE) segmentation is used to identify an initial set of breast structures. The DWCE segmentation employs an adaptive filter to enhance the local contrast and accentuate mammographic structures on the image. After contrast enhancement, Laplacian-Gaussian edge detection is applied and all enclosed objects are filled to produce a set of detected structures for the image. These objects are then used as starting locations for a clustering-based region-growing algorithm. First, an initial set of seed objects are determined by identifying all local maxima in the original gray-scale image. K-means clustering is then applied to the background-corrected regions of interest (ROIs) defined by each object. Since the DWCE segmentation and growing do not differentiate masses from normal tissues, a large number of objects are usually detected in each mammogram. A set of features is extracted from each detected object and used to differentiate between masses and normal breast structures. A classifier employing 11 morphological features is initially used to eliminate objects that had shapes significantly different from breast masses. Texture features are then computed for all remaining structures and used with a linear classifier as a final arbiter between potential masses and normal structures. The performance of this mass detection program on a training set of 253 mammograms achieved a 81% sensitivity at a false-positive rate of 2.1 per image and a 85% sensitivity at

a false-positive rate of 2.1 per image for malignant masses. The free response receiver operating characteristic (FROC) curve of the detection at all thresholds is shown in Fig. 2. Three different methods of scoring the detection are shown: the single threshold method, and methods using the hybrid 1 and hybrid 2 classifiers. The single threshold classifiers simply applies a single global threshold to all detected structures. The hybrid 1 classifiers normalizes the scoring between images by rescaling the maximum and minimum score within each image to 1 and 0, respectively. A single threshold between 1 and 0 is then applied. The hybrid 2 classifier keeps at most the detected objects with the highest 3 scores. A single threshold, without any rescaling, is then applied to this reduced set of objects.

We have evaluated the performance of our mass detection program with randomly selected test cases from patients with biopsy-proven masses. An experienced radiologist identified all the masses on each image and the locations of the mass on the digitized mammogram were stored in a truth file for scoring of the detection results. For a test data set containing 233 masses, the program achieved a sensitivity of about 82% at a false-positive rate of 2.1 per image for malignant masses (Fig. 3). If all malignant and benign masses were taken into account, the detection sensitivity was 73% at a false-positive rate of 2.2 per image (Fig. 4). For a test data set of 100 mammograms collected at the Georgetown University, the program achieved a sensitivity of 74% at a false-positive rate of 2 per images (see Fig. 6 in Georgetown University report below). This performance is reasonable taking into account the fact that these cases are truly independent of the training data set. Furthermore, both the screen-film system and the film digitizer are different from those used for the training cases.

Detection of Microcalcifications

We have completed the integration of the microcalcification detection program into the automated CAD system. Some modifications of the microcalcification detection program have also been incorporated into the program to improve the detection accuracy. For this program, an input digitized mammogram is preprocessed with the same edge trimming and breast area segmentation program as for mass detection. The breast region of the digitized mammogram is processed with spatial filters to obtain the signal-enhanced and signal-suppressed images. A difference image is then obtained by subtracting the signal-suppressed image from the signal-enhanced image. Since the low-frequency structured background is similar in the two images, the difference image technique removes the slowly varying background from the difference image. An adaptive gray-level thresholding technique is then applied to the difference image in order to segment potential microcalcifications from the remaining noise background. The resulting threshold image contains groups of pixels with values above the threshold superimposed on a uniform background. Potential microcalcifications are identified in the threshold image using an area-thresholding criterion that eliminates random noise points with areas smaller than a preselected number of pixels. Additionally a convolution neural network trained to recognize true microcalcification patterns is used to reduce false positives. Finally a clustering criterion is used to identify microcalcification clusters containing more than a preselected number of detected microcalcifications within a predefined diameter. We have evaluated the performance of the microcalcification detection program with randomly selected cases containing biopsy-proven microcalcifications. With a test data set of 260 images that contained 77 malignant and 143 benign microcalcification clusters, the detection program achieved a sensitivity of 79% at a false-positive rate of 0.8 cluster per image. The FROC curve at different decision thresholds is plotted in Fig. 5.

(b) CADView workstation

Improvement of the CAD visualization system

The design and operation of the graphical user interface (GUI) have been discussed in details in last year's report. In this year, we have set up a PC workstation with a pentium III processor and a liquid crystal display (LCD) monitor in our off-line reading room for the CAD reading and recording of the radiologists' evaluation. This PC workstation is named "CADView." We have conducted preliminary clinical testing of the CADView for display of the computer detection results during film interpretation. Several experienced mammographers evaluated the results and the GUI. These radiologists are clinically oriented and they are not involved in the development of the GUI. The purpose of the evaluation is to assess the practicality of the GUI for radiologists who are not familiar with the CAD project and the use of computer outputs. Based on the radiologists' suggestions and comments, revisions are made to improve the user-friendliness of CADView for daily clinical use. For the purpose of the pilot clinical study, we have decided to record not only the action category but also the BI-RADS assessment by the radiologist before and after the computer detection results are displayed. The evaluation results are recorded in our database. After testing with the radiologists who are not familiar with CAD, many feedback protections have been implemented in order to avoid skipping of scoring or inappropriate reading sequence. These modifications ensure that complete reading results will be collected from every reader and every case to be read.

Training experiment

We have designed a small-scale training experiment to familiarize the radiologists with the CAD workstation and the pilot clinical study. A set of 15 screening cases was selected from our recent patient files. All cases included a current exam and an exam from the previous year that were read as normal. In addition, three cases with subtle abnormalities that were biopsied in the current year were also selected and randomly mixed with the normal cases. The films in the previous year were processed by the detection programs to obtain computer detection results. Four experienced radiologists were asked to read the previous exam without and with CAD in a setting similar to our planned pilot clinical study. After the reading and recording of their decision without and with CAD using the CADView workstation, the current-year exam was presented to the radiologists. The radiologists could thus learn the characteristics of the true-positive and false-positive detections by the computer. This study was important because it familiarized the radiologists with the performance of the computer. This would reduce the possibility that radiologists would be over-sensitized by the computer output and increased the call back rates.

(c) Implementation of CADView workstation at Georgetown University

We have installed a CADView workstation and a high speed UNIX AlphaStation at the Georgetown University for the pilot clinical study. The Alpha workstation and the CADView were set up in the same way as the systems at the University of Michigan. The operation of the CADView and the processing of the images on the AlphaStation are the same in both sites. This facilitates the maintenance and the upgrade of the system software. Our collaborators, Dr. Lo and Dr. Freedman at Georgetown University have tested the systems and performed evaluation on the detection programs using cases with biopsy-proven masses and calcifications. They have reported the details of the implementation and testing in the following sections.

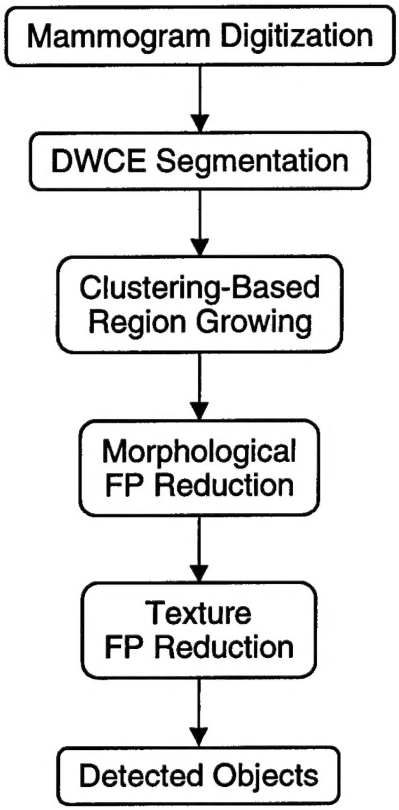


Fig. 1. Block diagram of the current mass segmentation method.

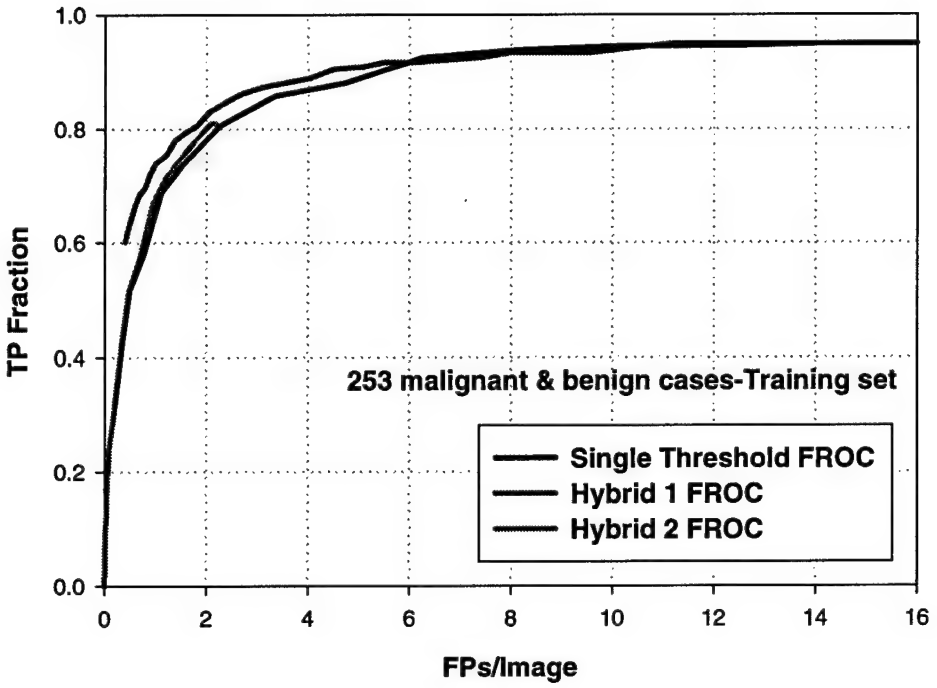


Fig. 2. The overall performance achieved by the mass detection program for a training set of 253 mammograms. The detection accuracy is represented by an FROC curve. Three different methods of scoring the detection are shown: the single threshold method, and methods using the hybrid 1 and hybrid 2 classifiers. The single threshold classifiers simply applies a single global threshold to all detected structures. The hybrid 1 classifiers normalizes the scoring between images by rescaling the maximum and minimum score within each image to 1 and 0, respectively. A single threshold between 1 and 0 is then applied. The hybrid 2 classifier keeps at most the detected objects with the highest 3 scores. A single threshold, without any rescaling, is then applied to this reduced set of objects. The same legend applies to Figs. 3, 4, and 6 below and will not be repeated there.

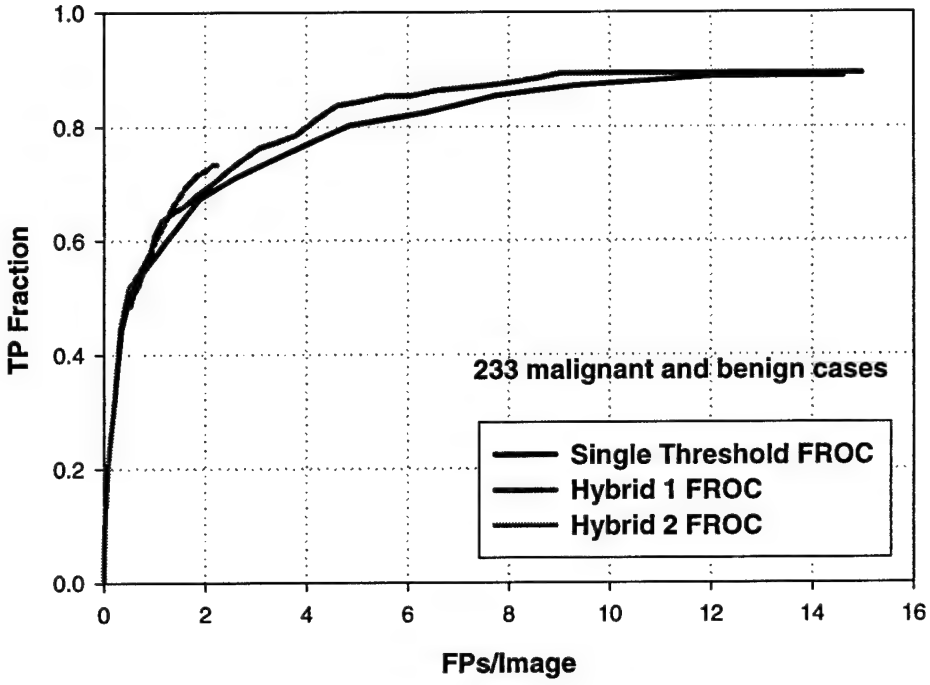


Fig. 3. The overall performance achieved by the mass detection program for a test data set of 233 masses.

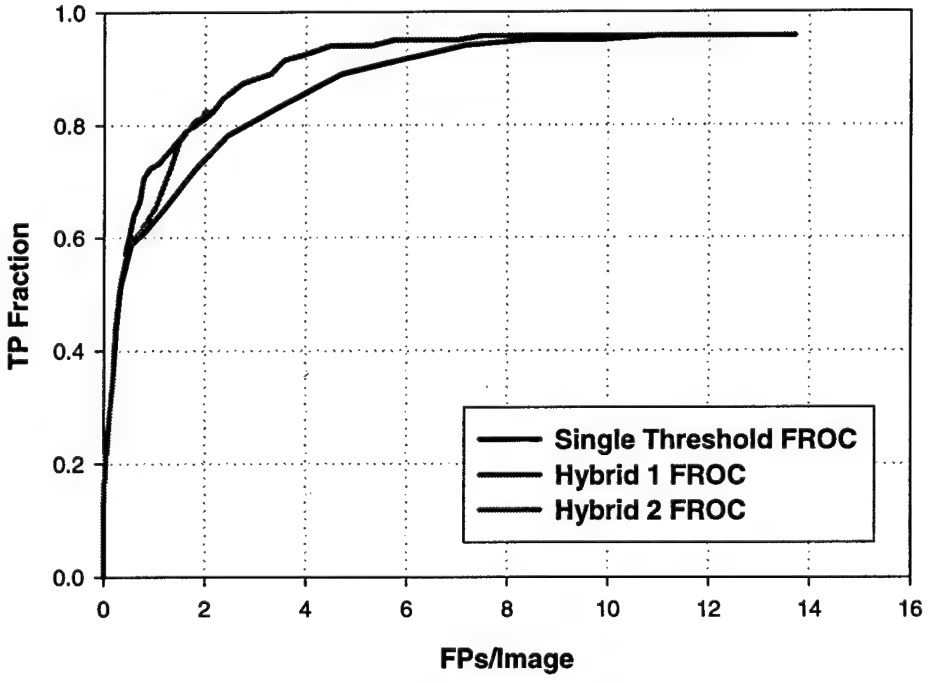


Fig. 4. The performance achieved by the mass detection program for a data set of 119 mammograms containing malignant masses.

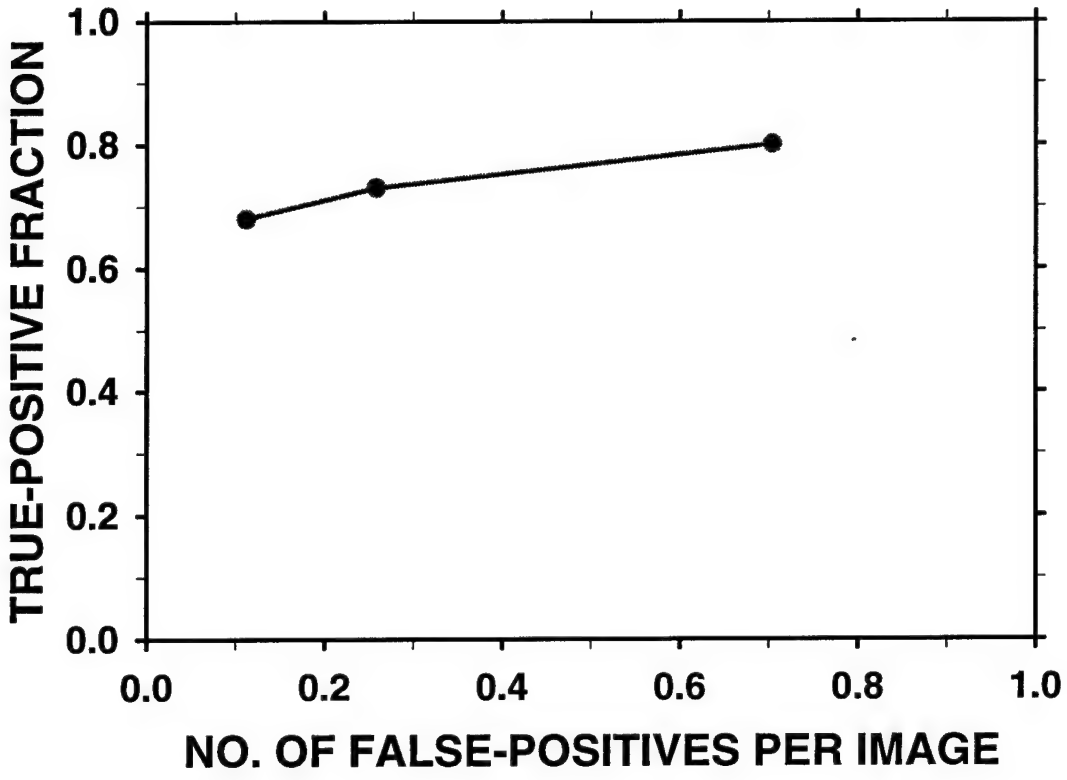


Fig. 5. The performance achieved by the microcalcification detection program for a test data set of 260 mammograms containing 77 malignant and 143 benign microcalcification clusters.

Georgetown University

Annual Report (9/23/98 – 9/22/99) to USAMRMC through University of Michigan

(a) System Implementation for Computer-Aided Detection Clinical Trial at Georgetown University

Although the research team at the ISIS Center, Georgetown University Medical Center has independently developed a computer-aided diagnosis system for the detection of clustered microcalcifications and masses on mammograms in the past 8 years, we decided to use the CAD system developed by the research team led by Dr. Heang-Ping Chan (the PI of this project) at the University of Michigan to prepare and perform the clinical trial. Initially we attempted to install the detection programs on a SUN workstation. However, despite efforts to modify the programs to accommodate both systems, we found that there were a number of system incompatibility problems between an COMPAQ/DEC AlphaStation and a SUN workstation, which may hinder management and analysis of the data collected from the study. Furthermore, the computation speed of the SUN workstation is too slow to meet the demand of processing the screening mammograms. In order to avoid system incomparability, we purchased a high-speed COMPAQ/DEC XP1000 workstation and a PC workstation with our own fund to conduct the study. Dr. Nick Petrick and Dr. Lubomir Hadjiyski from the University of Michigan loaded and tested their software in the XP1000 and the PC workstation, respectively. Dr. Ben Lo and Mr. Andrzej Delegacz also worked at the Georgetown side to perform various tests suggested by Dr. Heang-Ping Chan. The system and software installation as well as initial tests using existing digitized mammograms have now been completed. Some of the testing results are summarized in Fig. 5 and Table 1 in Section (b) below.

Georgetown team has begun to perform the clinical trial in the breast imaging division. Currently, Drs. Freedman and Makariou have used the system to perform their routine clinical practice. We put a Lumisys film digitizer (Model Lumiscan 150) hosted by a SUN SPARC 10 workstation at the Breast Imaging Division, Radiology Department, Georgetown University Medical Center. The data flow is chained through a 3-step processing.

Step 1: A mammogram is digitized at the SUN/Lumiscan workstation. Patient information, including ID, age, side, view (CC or MLO) and examination date, is recorded during the digitization and entered into CAD patient/film database (part of the CADView system) on the PC computer. Each mammogram is digitized at 100 micron resolution. The image files are stored at a designated directory at the SUN workstation hard disk. The image files are also transferred for further processing to the XP1000 workstation at the ISIS Center via a high-speed Ethernet connection.

Step 2: A control program running on the XP1000 workstation continuously searches for new images being transferred from the SUN/Lumiscan workstation. When a new image appears, this control program initiates the execution of the program to detect the mass and clustered microcalcifications on that image and stores the detection results in appropriate directories.

Step 3: On the PC workstation, a CADView program, designed and implemented by Drs. Lubomir Hadjiyski and Heang-Ping Chan and their co-workers, is used as the user interface to review and analyze the results of the mass and microcalcification detection. The CADView program uses the automated procedure to download the result images from the XP1000 workstation on an on-demand basis. The radiologist uses the patient ID number to retrieve patient information from

the database (updated in step 2), including information on images to be displayed, and show it on the screen. If the requested images are not available locally, the program establishes the FTP session with the XP1000 workstation and downloads those image files to its working directory on the PC workstation. The radiologist can then perform the clinical evaluation of the patient films. The program, among others, allows the radiologist to mark the location of any visible masses and/or microcalcifications on the images, along with his/her action rating. The results of the radiologist's review and evaluation are stored in the database.

(b) Initial Tests on mammographic cases

We used 63 mammographic cases containing 268 mammograms, which were collected as a part of Dr. Matthew T. Freedman's teaching files with radiographic reading reports, to test the performance of the installed system. Dr. Letitial Clark, a MQSA mammographer at GUMC, was invited to identify the abnormality on the mammograms based on the previously recorded reports without biopsy results. Since the original films were back to the film library, we use the monitor to display the image. The display software with window/level and zoom-in/zoom-out functions was used for viewing the mammograms. Dr. Clark read each radiological report prior to identifying the location of the associated mammograms that an abnormality was reported. The display program is also equipped with a user interface that can record the location identified by the radiologist. The window/level function was always employed while looking at clustered microcalcifications. In this set of mammograms, 7 out of 63 cases have no clinically significant signs of breast cancer: one case contains scattered calcifications, one case collected from a follow-up mammography (the lesion of which has been removed from the breast), and five cases show insignificant asymmetrical density. 43 cases were identified to have abnormalities associated with masses or asymmetrical density. 19 cases were identified abnormalities associated with clustered calcifications. 9 cases have both masses and clustered calcifications. Dr. Clark identified a total of 106 masses on 268 mammograms, of which 100 are primary masses.

The 100 mammograms that contained masses were processed with the mass detection program on the AlphaStation. In the preprocessing stage, the program trimmed the blank edges around the film and segment the breast area from the mammogram. The trimming did not work well in nine of the 100 images. However the detection still seemed to work relatively well even when the trimming had some problems. In Table 1, we reported both the overall results and the results without the bad trimming cases and also compared the detection with the training and test results from the University of Michigan. Fig. 6 shows the FROC curve for the entire range of detection thresholds. The true-positive fraction (TPF) and the number of false-positive per image (FP/Img) were very similar to those obtained at the University of Michigan. The detection program achieved a TPF of 74% at an FP rate of about 2 per image. Since the biopsy results of the masses is not known, we cannot report results of the malignant and benign masses separately.

Table 1. Summary of detection results of the mass detection program.

Cases	Mass Types	No. Images	No. Masses	No. Detected	TPF	FPs/Img
UM Training	Malig and Benign	253	253	205	81.0	2.1
UM Training	Benign	125	125	96	76.8	2.2
UM Training	Malignant	128	128	109	85.2	2.1
UM Test	Malig and Benign	233	233	171	73.4	2.2
UM Test	Benign	114	114	73	64.0	2.3
UM Test	Malignant	119	119	98	82.4	2.1
Georgetown Test	Primary masses	100	100	75	75.0	2.0
Georgetown Test	Primary masses (bad trimming cases removed)	91	91	69	75.8	2.0
Georgetown Test	All masses	100	106	78	73.6	2.0
Georgetown Test	All masses (bad trimming cases removed)	91	97	72	74.2	2.0

(c) On Global Segmentation of Large Regions of Interest – A unified theory

In the past few years, there has been a large number of publications that discussed about the delineation of a large region of interest (ROI) in the field of digital radiography. The topics include (1) breast and glandular areas in mammography, (2) lung field and ribs in chest radiography, (3) lung and heart in chest CT image sequence, and (4) liver in abdominal CT images etc. Although this project aims to the technical development and clinical evaluation for computer-aided detection of breast cancer and digital compression in mammography, we would like to report a technical finding regarding global segmentation of a large area which is applied to the segmentation of breast and glandular tissues.

This finding was initiated by training and analyzing the convolution neural network (CNN) [Lo 1995]. Although the CNN would take a long training iteration, the internal kernels function as convolution filters. Technically, these kernels can be combined into a single kernel. One can also deconvolve the original input image by the output image to obtain the filter that is equivalent to the CNN process. Although the resulting filters obtained from the deconvolution process may not be identical for every pairs of input and output images, our experiment indicates that they are quite close as far as global segmentation of a large area is concerned for a specific type of images. To be exact, we found that a single linear filter can be found for each type of image segmentation mentioned above. In addition, all linear filters discovered so far belong to a single set of filter family. This discovery allows us to construct a unified theory as follow.

Theoretically speaking, the frequency band associated with large area without detail structures should be predominated by low frequency. The results, obtained by composing the CNN kernel discussed above, prove this fact. We, therefore, hypothesize that the criteria of low frequency filters for global segmentation should contain (i) significant amounts of low frequency components, (ii) very few or no high frequency components, (iii) no band frequencies associated with the structures that were intended to be removed, and (iv) a low (or zero) mean coefficients (i.e., $\sum_{x,y} t(x,y) \approx 0$.) Although, a great deal of low frequency filter banks are available, there are three types of known filters commonly used in digital signal processing: (a) uniform low-pass filters in frequency domain, (b) local mean value operators (i.e., uniform column filter) in spatial domain, and (c) Gaussian shape filters. Since main frequency components in (a) can be approximately described through (b) and (c), we can assume that the filter to be constructed is composed of (b) and (c) components in this study.

$$t(x,y) = m(x,y) + G(x,y) \qquad \ldots(1)$$

where

$$m(x,y) = \begin{cases} w_1 & for \ \ (x^2+y^2)^{1/2} \le r \\ 0 & for \ \ (x^2+y^2)^{1/2} > r \end{cases}, \quad G(x,y) = w_2 \exp\left((x^2+y^2)/2\sigma^2\right)$$

and r is the cut-off range of the uniform column filter. In addition, the uniform column filter possesses a property of $\pi r^2 w_1 = 1$. Substitute the components of the m(x,y) and G(x,y) filter in eq. (1), we have

$$t(x,y) = \begin{cases} w_1 + w_2 \exp\left(-(x^2+y^2)/2\sigma^2\right) & for \ \ (x^2+y^2)^{1/2} \le r \\ w_2 \exp\left(-(x^2+y^2)/2\sigma^2\right) & for \ \ (x^2+y^2)^{1/2} > r \end{cases} \qquad \ldots(2)$$

We further constraint that the composed filter should be a filter with zero mean coefficients (i.e., $\sum_{x,y} t(x,y) = 0$), which is a common requirement when designing an edge enhancement filter. With this constraint, a logical solution in eq. (1) is

$$\pi r^2 w_1 = -w_2 \iint_{x,y} \exp\left((x^2+y^2)/2\sigma^2\right) = 1.$$

Hence, $w_1 = \frac{1}{\pi r^2}$ and $w_2 = \frac{-1}{2\pi\sigma^2}$. Substituting them into eq. (2), we have

$$t(x,y) = \begin{cases} \frac{1}{\pi r^2} - \frac{1}{2\pi\sigma^2}\exp\left(-(x^2+y^2)/2\sigma^2\right) & for \quad (x^2+y^2)^{1/2} \le r \\ \frac{-1}{2\pi\sigma^2}\exp\left(-(x^2+y^2)/2\sigma^2\right) & for \quad (x^2+y^2)^{1/2} > r \end{cases} \quad \ldots(3)$$

Our initial experience indicates that 2r should be about the size of the largest structure to be removed. The digital form of the composed filter can be implemented below:

$$t(x,y) = \begin{cases} \frac{1}{A_m} - \frac{1}{A_G}\exp\left(-(x^2+y^2)/2\sigma^2\right) & for \quad (x^2+y^2)^{1/2} \le r \\ \frac{-1}{A_G}\exp\left(-(x^2+y^2)/2\sigma^2\right) & for \quad (x^2+y^2)^{1/2} > r \end{cases} \quad \ldots(4)$$

where $A_m = \sum_{x,y} 1 \quad for \quad (x^2+y^2)^{1/2} \le r$ and $A_G = \sum_{x,y} \exp\left((x^2+y^2)/2\sigma^2\right)$

Determination of the filter parameters

The sample mammograms, used in our study, were digitized at 0.03584 cm per pixel. For the segmentation of breast area, we determined that 2r is 40 pixels (1.433 cm) for the flat column filter and 2σ is 100 pixels (3.584 cm) in the Gaussian filter (Fig. 7(A)). Hence, A_m = 1,245 and A_G = 15,708. The sample chest radiographs used in our study were digitized at 0.07 cm per pixel. For the segmentation of lung regions, we determined that 2r is 18 pixels (1.26 cm) for the flat column filter and 2σ is 74 pixels (5.18 cm) in the Gaussian filter (Fig. 7(B)). These parameters were determined with consideration of trimming ribs. Hence, A_m = 249 and A_G = 8,602. The same filter also applied to CT chest images for segmentation of lungs. Sample images and their results are shown in Figs. 8, 9, 10. These results indicate that lung segmentation has been very successful. However, the filter parameters set above for the breast has not yet been optimized. The filtered breast area is somewhat underestimated. We will test more breast images and report a better set of parameters using this unified approach.

(d) Integer Wavelet Compression in Mammography

Collected database

A total of 530 sets of mammograms were collected and digitized by the SUN/Lumiscan workstation. Each case contains 4 mammograms (2 sides and 2 views). Some cases (less than 10%)

contain 1 side 2 views. A total of 310 mammograms (about 50% MLO and 50% CC views) randomly selected from the database have been identified as a subset for this compression study.

Initial visual study using the decompressed mammograms

We have tested 20 mammograms consisting of a variety of breast parenchymal patterns. Each mammogram was compressed at 0.3 bit/pixel, 0.2 bit/pixel, 0.175 bit/pixel, 0.15 bit/pixel, 0.125 bit/pixel, and 0.1 bit/pixel initially. We also found that an average of 700 patches (ranging from 400 to 1200) were identified by the CAD program. This implies that an average of 700x10x10 pixels x 8 bit must be added to the compressed file for lossless compression at the patches. In other words, approximately 0.1 bit/pixel shall add to the bit rate. Our initial visual inspection indicates that no visual degradation can be observed with 0.3+0.1 bit/pixel compression for all mammograms. However, some subtle (blur) artifacts can be observed with 0.2+0.1 bit/pixel in large mammograms. The artifacts are barely observed on small breasts with 0.15+0.1 bit/pixel. This study serves as a guide for the compression ratio to be tested in the full-sized study.

Execution of the compression program and planning for the comparison study

Based on the initial study, we have decided to compress each mammogram with bit rates at 0.3+0.1 bit/pixel and 0.15+0.1 bit/pixel. The compressed files have been stored in SUN tapes and are ready for the subjective comparison study. We will continue to add more mammograms in this compression data set. Our goal is to obtain 600 mammograms, each from a single case, in the first quarter of 2000.

We plan to use our own fund to purchase a mammography workstation ($42,000) developed by Imaging Smith. Dr. Jerry Gaskil of Imaging Smith has shown Drs. Matthew T. Freedman (clinical director of this project at GUMC) and Dr. Heang-Ping Chan the speed and capability of workstation. The system, possessing a dual-CPU PC, a large disk space, a high-speed graphic board, and two high-resolution monitors, has been used as the main display workstation for the digital mammography project at the Naval Hospital (Bethesda, Maryland). We are negotiating with Imaging Smith for the system functionality and upgrade for NT window 2000 and plan to purchase a machine in the spring of 2000. After the system is installed at Georgetown, we will load the original and compressed files and perform the subjective comparison study.

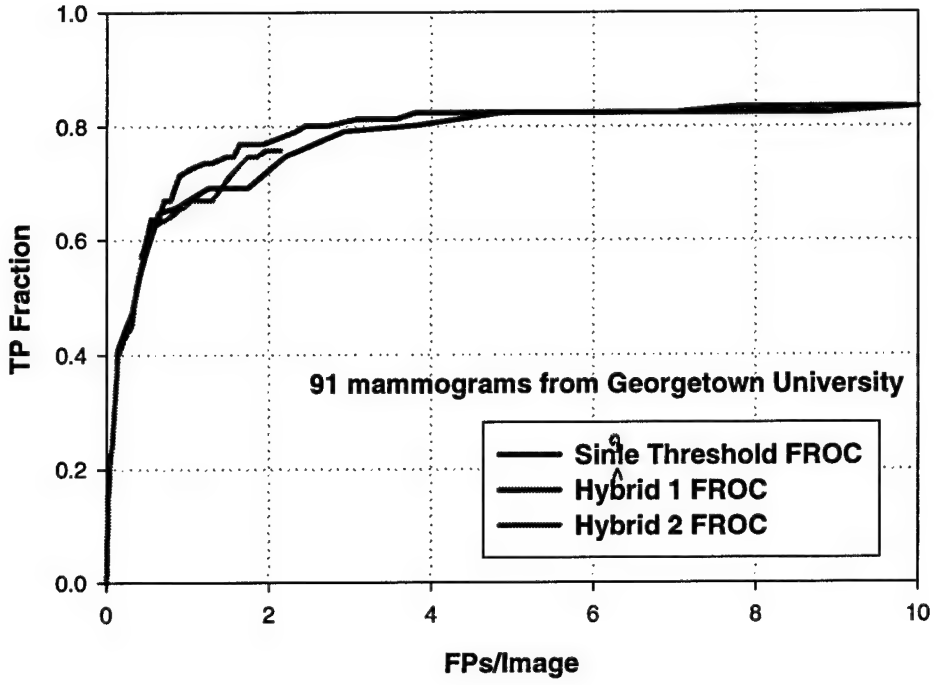
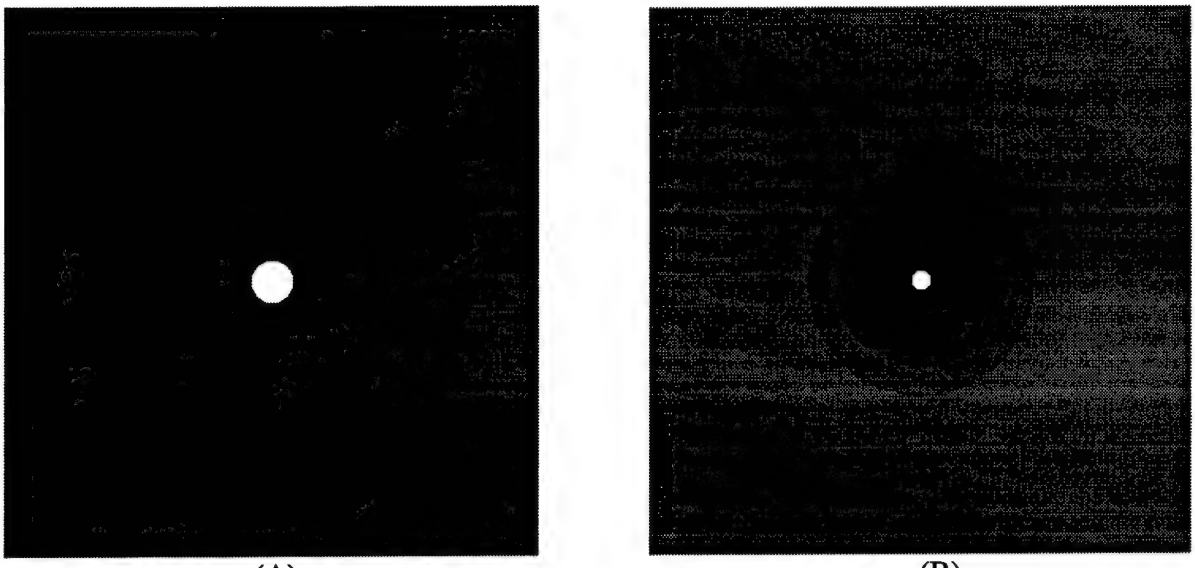
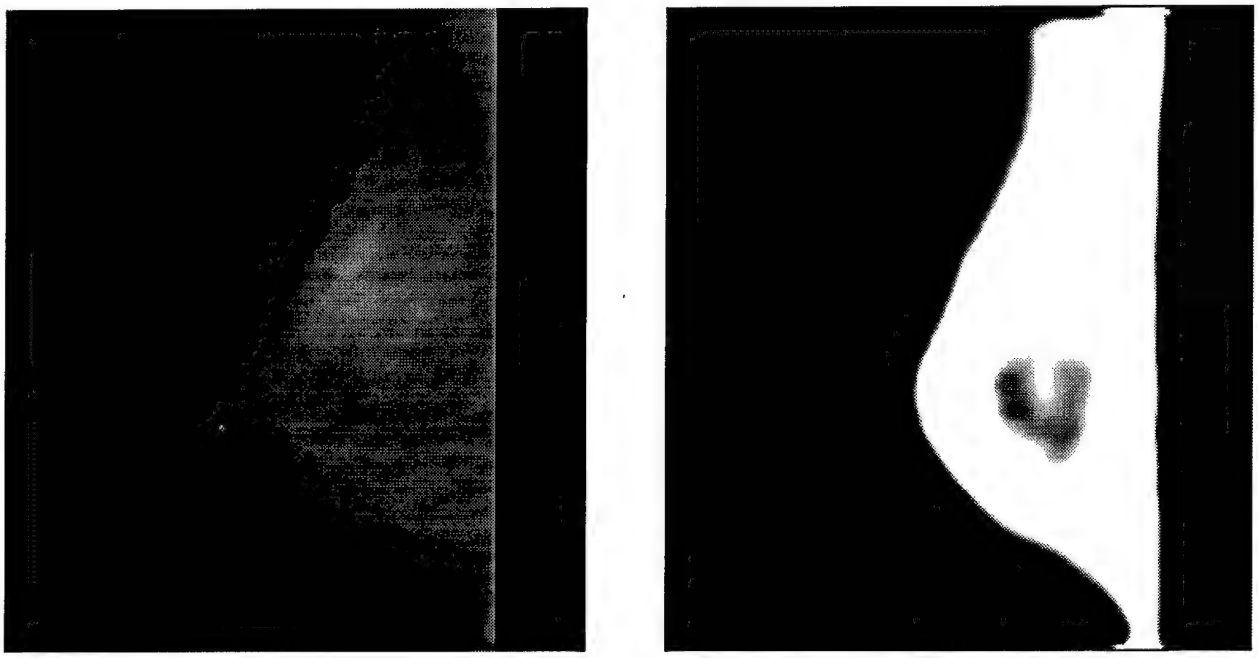


Fig. 6. The overall performance achieved by the mass detection program for a data set of 91 mammograms containing masses from Georgetown University.



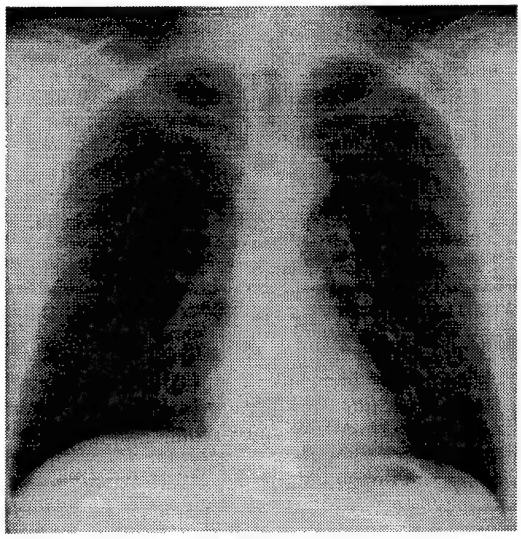
(A) (B)

Fig. 7. Filter designed for the segmentation of breast area in mammography (A) and lung in chest radiography (B). Note that both filtered images were scaled to 255 for highest values for the display purpose which means the scaling factor is ($255A_m$).

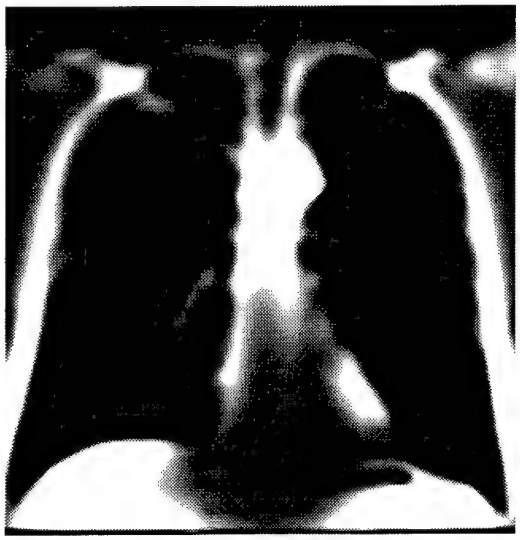


(A) (B)

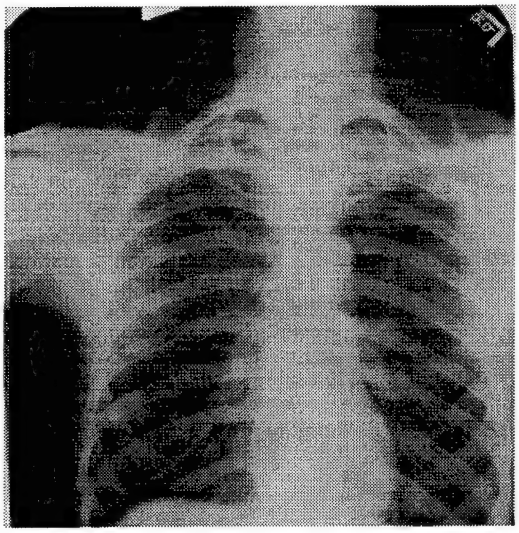
Fig. 8. The original breast image and its segmented image filtered by Fig. 7(A).



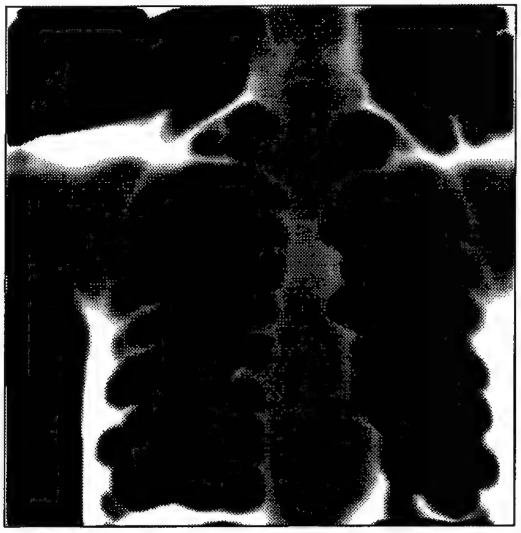
(A1)



(B1)

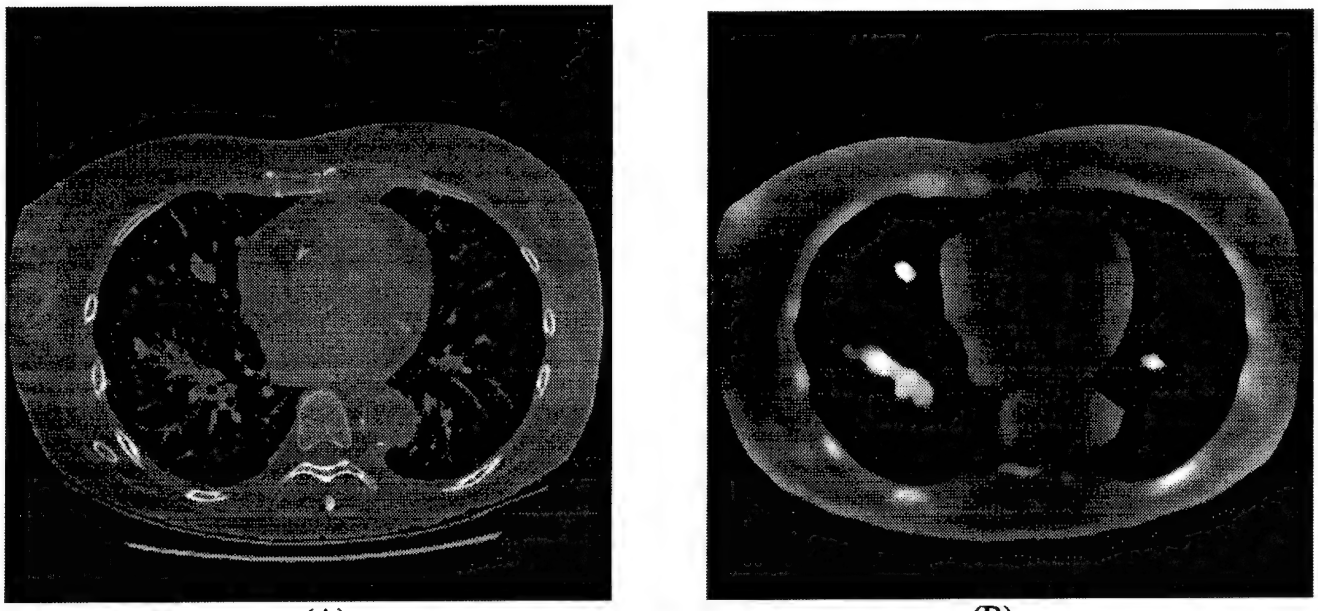


(A2)



(B2)

Fig. 9. The original chest images (A1 & 2) and their segmented images (B1 & 2) filtered by Fig. 7(B).



(A) (B)

Fig. 10. The original CT chest image (A) and its segmented image (B) filtered by Fig. 7(B).

University of Iowa

Development of Methods for Analyzing Pilot Clinical Trial Data

We have been testing the applicability of the Dorfman, Berbaum, Metz (1992) multireader, multipatient (MRMP) methodology for analyzing receiver operating characteristic (ROC) data from the clinical trial. The CAD workstation implements the American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) final categories. In clinical trials, these categories are action categories and have implications for patient care. The category "negative" translates into one year followup, "probably benign finding" translates into the course of action "short interval followup suggested," "suspicious abnormality" translates into the course of action "biopsy should be considered," and "highly suggestive of malignancy" translates into "appropriate action should be taken". Some diagnostic imaging systems may lead to more conservative or liberal actions than others. We plan to estimate decision thresholds associated with the action categories using proper ROC analysis (Dorfman, Berbaum, Metz et al., 1997). Proper ROC analysis is essential for this pilot clinical trial because of the paucity of cancers.

We have tested the Dorfman/Berbaum/Metz (DBM) methodology with a comprehensive series of computer simulations on factorial experimental design (Dorfman, Berbaum, Lenth et al., 1998). The results suggest that the DBM method provides trustworthy alpha levels with discrete ratings when ROC area is not too large, and case and reader sample sizes are not too small. In other situations, the test tends to be somewhat conservative or very slightly liberal. We have also tested the DBM methodology with a comprehensive series of computer simulations on split plot experimental design (Dorfman, Berbaum, Lenth et al., 1999). Our Monte Carlo simulations show that the DBM multireader methodology can be validly extended to the split plot design where readers interpret imaging studies of different patients in CAD vs no CAD conditions. Both of these validation studies used a balanced design, which is appropriate for laboratory studies, but perhaps not for clinical trials.

We have implemented the DBM methodology for unbalanced designs in the event that different readers finish with a different numbers of imaging studies read in CAD and no CAD conditions. To achieve this goal, we have distributed the function of MRMP so that we can perform statistical analyses with SAS. We used a dynamic link library (DLL) containing RSCORE and a subroutine called JACKKNIFE that computes pseudovalues from the raw categorical rating data. The pseudovalues are submitted to SAS for statistical analysis (Littell et al. 1996). This means that all of the mixed-model programs of SAS are available for multireader multipatient ROC analysis. This is a very flexible procedure for analyzing a wide variety of multireader multipatient ROC data, and is ideally suited for analyzing data from clinical trials that have the goal of comparing diagnostic modalities.

(7) Key Research Accomplishments

- Implementation of CAD workstation and the associated CAD software at the University of Michigan and the Georgetown University.
- Implementation and testing of the CAD result visualization software and graphical user interface (CADView) at the University of Michigan and the Georgetown University.
- Testing of mass detection program with 330 mammograms and evaluate the detection accuracy software at the University of Michigan and the Georgetown University.
- Testing of microcalcification detection program with 260 mammograms and evaluate the detection accuracy.
- Preliminary evaluation of effects of CAD on mammographic interpretation by experienced radiologists.
- Collection of database for observer experiment to evaluate effects of image compression on mammographic image quality.
- Development of unified approach for image segmentation by Georgetown University.
- University of Iowa - Performance of comprehensive Monte Carlo simulation study of multi-reader, multi-patient method for analysis of unbalanced design of ROC studies with BI-RADS scoring, in preparation of analysis of data collected from the pilot clinical study.

(8) Reportable Outcomes

Publications

Journal Articles

1. Chan HP, Sahiner B, Helvie MA, Petrick N, Roubidoux MA, Wilson TE, Adler DD, Paramagul C, Newman JS, Sanjay-Gopal S. Improvement of radiologists' characterization of mammographic masses by computer-aided diagnosis: an ROC Study. Radiology 1999; 212: 817-827.

2. Petrick N, Chan HP, Sahiner B, Helvie MA, Goodsitt MM. Combined adaptive enhancement and object-based region growing for automated detection of masses on mammograms. Medical Physics 1999; 26: 1642-1654.

Articles Accepted for Publication:

1. Sanjay-Gopal S, Chan HP, Wilson TE, Helvie MA, Petrick N, Sahiner B. A regional registration technique for automated interval change analysis of breast lesions on mammograms. Medical Physics 1999; 26 (in press, December).

2. Chan HP, Sahiner B, Wagner RF, Petrick N. Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers. Medical Physics 1999; 26 (in press, December.

3. Sahner B, Chan HP, Petrick N, Wagner RF, Hadjiiski LM. Feature selection and classifier performance in computer-aided diagnosis: the effect of finite sample size. Medical Physics.

Conference Proceedings

1. Chan HP, Helvie MA, Petrick N, Sahiner B, Roubidoux MA, Wilson TE, Joynt LK, Hadjiiski LM, Paramagul C, Adler DD, Goodsitt MM. Digital Mammography: observer performance study of the effects of pixel size on radiologists' characterization of malignant and benign microcalcifications. Proc. SPIE 1999; 3659: 394-397.

2. Sahiner B, Chan HP, Petrick N, Wagner RF, Hadjiiski LM. Stepwise linear discriminant analysis in computer-aided diagnosis: the effect of finite sample size. Proc. SPIE 1999; 3661: 499-510.

3. Hadjiiski LM, Sahiner B, Chan HP, Petrick N, Helvie MA. Hybrid unsupervised-supervised approach for computerized classification of malignant and benign masses on mammograms. Proc. SPIE 1999; 3661: 464-473.

4. Dorfman DD, Berbaum KS, Lenth RV, Chen Y-F. Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: Split plot experimental design. Proc. SPIE 1999; 3663: 91-99.

Abstracts, Presentations, Scientific Exhibits

1. Chan HP, Helvie MA, Petrick N, Sahiner B, Roubidoux MA, Wilson TE, Joynt LK, Hadjiiski LM, Paramagul C, Adler DD, Goodsitt MM. Digital Mammography: observer performance study of the effects of pixel size on radiologists' characterization of malignant and benign microcalcifications. Presented at the SPIE International Symposium on Medical Imaging, San Diego, CA, February 20-26, 1999.

2. Sahiner B, Chan HP, Petrick N, Wagner RF, Hadjiiski LM. The effects of sample size on feature selection in computer-aided diagnosis. Presented at the SPIE International Symposium on Medical Imaging, San Diego, CA, February 20-26, 1999.

3. Hadjiiski LM, Sahiner B, Chan HP, Petrick N, Helvie MA. Hybrid unsupervised-supervised approach for computerized classification of malignant and benign masses on mammograms. Presented at the SPIE International Symposium on Medical Imaging, San Diego, CA, February 20-26, 1999.

4. Petrick N, Chan HP, Goodsitt MM, Sahiner B, Hadjiiski LM. Digital mammographic imaging using microlens focusing: Estimates of light collection and x-ray utilization. Presented at the SPIE International Symposium on Medical Imaging, San Diego, CA, February 20-26, 1999.

5. Wagner RF, Chan HP, Sahiner B, Petrick N, Mossoba JT. Components of variance in ROC analysis of CADx classifier performance. II: Applications of the bootstrap. Presented at the SPIE International Symposium on Medical Imaging, San Diego, CA, February 20-26, 1999.

6. Chan HP, Hadjiiski LM, Petrick N, Helvie MA, Roubidoux MA, Sahiner B. Performance evaluation of an automated microcalcification detection system. Accepted for presentation at the 41st Annual Meeting of the American Association of Physicists in Medicine. Nashville, Tennessee, July 25-29, 1999.

7. Chan HP, Sahiner B, Helvie MA, Petrick N, Hadjiiski LM, Roubidoux MA. Computer-aided breast cancer diagnosis: Comparison of computerized classification with radiologists' performance. Accepted for presentation at the 85th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Nov. 28-Dec. 3, 1999, Chicago, Illinois.

8. Hadjiiski LM, Chan HP, Sahiner B, Petrick N, Helvie MA, Sanjay-Gopal S. Automated identification of breast lesions in temporal pairs of mammograms for interval change analysis. Accepted for presentation at the 85th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Nov. 28-Dec. 3, 1999, Chicago, Illinois.

9. Sahiner B, Chan HP, LeCarpentier GL, Petrick N, Roubidoux MA, Carson PL. Computerized characterization of solid breast masses using three-dimensional ultrasound images. Accepted for presentation at the 85th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Nov. 28-Dec. 3, 1999, Chicago, Illinois.

10. Petrick N, Chan HP, Sahiner B, Helvie MA, Paquerault S. Evaluation of an automated computer-aided diagnosis system for the detection of masses on prior mammograms. Accepted for poster presentation at the SPIE International Symposium on Medical Imaging, San Diego, CA, February 12-18, 2000.

11. Zhou C, Chan HP, Petrick N, Sahiner B, Helvie MA, Roubidoux MA, Hadjiiski LM, Goodsitt MM. Computerized image analysis: Estimation of breast density on mammograms. Accepted for poster presentation at the SPIE International Symposium on Medical Imaging, San Diego, CA, February 12-18, 2000.

12. Hadjiiski LM, Chan HP, Sahiner B, Petrick N, Helvie MA, Paquerault S, Zhou C. Interval change analysis in temporal pairs of mammograms using a local affine transformation. Accepted for poster presentation at the SPIE International Symposium on Medical Imaging, San Diego, CA, February 12-18, 2000.

(9) Conclusions

We have performed extensive evaluation of the computer detection programs and the GUI this year. The mass detection program has been evaluated with over 300 mammograms at the University of Michigan and the Georgetown University and the test performance of the program on these unknown cases was found to be about 73% at a false-positive rate of 2 per image. More importantly, the detection sensitivity of malignant masses was 82% at about 2 false-positive per image. The microcalcification detection program was evaluated with 260 mammograms and the program achieved a detection sensitivity of 79% at about 1 false-positive per image. This performance accuracy is at a reasonable level considering the fact that the test mammograms are independent of the training cases. In a small-scale reading experiment simulating the pilot CAD reading of screening mammograms by four experienced mammography radiologists, we found that the CAD could improve the detection of cancer cases, but there might be a very small increase in the call-back rate. We expect that the pilot clinical study will provide information if the increase is statistically significant.

Two CADView workstations have been implemented at the University of Michigan and the Georgetown University. The pilot clinical study in our off-line screening mammography clinics has begun and will collect data for the analysis of the effects of CAD on radiologists' reading.

The CAD-guided image compression project is progressing as planned. The compression technique has been evaluated in a small data set described in the GU report last year. A large data set has been assembled and the preparation for the observer evaluation study has been completed. The subjective image quality comparison study is planned to start early next year.

Because of the change in the strategy for the CAD workstation development and the addition of the mass detection program, as described in the previous reports, as well as the incompatibility of different workstations and operating systems, there is a delay in starting the pilot clinical study. We have requested and obtained approval for a no-cost-time-extension of one year to make up for part of the work.

(10) References

National Center for Health Statistics. Vital statistics of the United States, 1987. Vol. 2. Mortality. Part A, DHHS Publication no. (PHS) 90-1101 (Government Printing Office, Washington, D.C., 1990).

Baines CJ, Miller AB, Wall C and al e. Sensitivity and specificity of first screen mammography in the Canadian National Breast Screening Study: A preliminary report from five centers. Radiology 1986; 160: 295-298.

Bassett LW, Bunnell DH, Jahanshahi R, Gold RH, Arndt RD and Linsman J. Breast cancer detection: one versus two views. Radiology 1987; 165: 95-97.

Bird RE, Wallace TW and Yankaskas BC. Analysis of cancers missed at screening mammography. Radiology 1992; 184: 613-617.

Boring CC, Squires TS, Tong T and Montgomery S. Cancer statistics 1994. CA-A Cancer Journal for Clinicians 1994; 44: 7-26.

Byrne C, Smart CR, Cherk C and Hartmann WH. Survival advantage differences by age: evaluation of the extended follow-up of the Breast Cancer Detection Demonstration Project. Cancer 1994; 74: 301-310.

Chan HP, Doi K, Vyborny CJ, Schmidt RA, Metz CE, Lam KL, Ogura T, Wu Y and MacMahon H. Improvement in radiologists' detection of clustered microcalcifications on mammograms. The potential of computer-aided diagnosis. Invest Radiol 1990; 25: 1102-1110.

Curpen BN, Sickles EA, Sollitto RA and al. e. The comparative value of mammographic screening for women 40-49 years old versus women 50-59 years old. AJR 1995; 164: 1099-1103.

Dorfman DD, Berbaum KS, Metz CE, Lenth RV, Hanley JA, Abu-Dagga H. Proper receiver operating characteristic analysis: the bigamma model. Acad Radiol 1997; 4:138-149.

Dorfman DD, Berbaum KS, Lenth RV, Chen Y-F, Donaghy BA. Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: factorial experimental design. Acad Radiol 1998;5:591-602.

Dorfman DD, Berbaum KS, Lenth RV, Chen Y-F. Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: split plot experimental design. Proc. SPIE 1999;3663:91-99.

Feig SA and Hendrick RE, *Risk, Benefit, and Controversies in Mammographic Screening. In: Syllabus: A categorical Course in Physics Technical Aspects of Breast Imaging*, A. G. Haus and M. J. Yaffe, (Radiological Society of North America, Inc, Oak Brook, IL, 1993).

Freedman MT, Lo SCB, Artz, ST, Lau I, Mun SK, Classification of false positive findings on computer aided detection of breast microcalcifications. Proc. SPIE. 1997; 3034: 853-859.

Harris JR, Lippman ME, Veronesi U and Willett W. Breast Cancer. N Engl J Med 1992; 327: 319-328.

Harvey JA, Fajardo LL and Innis CA. Previous mammograms in patients with impalpable breast carcinomas: Retrospective vs blinded interpretation. AJR 1993; 161: 1167-1172.

Haug PJ, Tocino IM, Clayton PD and Bair TL. Automated management of screening and diagnostic mammography. Radiology 1987; 164: 747-752.

Hillman BJ, Fajardo LL, Hunter TB and al e. Mammogram interpretation by physician assistants. AJR 1987; 149: 907-911.

Kalisher L. Factors influencing false negative rates in xeromammography. Radiology 1979; 133: 297-301.

Kegelmeyer WP, Pruneda JM, Bourland PD, Hillis A, Riggs MW and Nipper ML. Computer-aided mammographic screening for spiculated lesions. Radiology 1994; 191: 331-337.

Littell RC, Milliken GA, Stroup WW, Wolfinger RD. SAS System for Mixed Models. Cary, NC: SAS Institute Inc. 1996.

Li, H, Lo SC, Wang Y, Freedman MT, and Mun SK: Mammographic mass Detection by Stochastic Modeling and a Multi-Module Neural Network, Proc. SPIE. 1997; 3034: 480-490.

Lo SCB, Chan HP, Lin JS, et al. Artificial convolution neural network for medical image pattern recognition. Neural Networks. 1995. 8:1201-1214.

Lo SCB, Xuan J, Li H, Wang YJ, Freedman MT, and Mun SK, Dyadic decomposition: A unified perspective on predictive, subband, and wavelet transforms," Proc. SPIE. 1997; 3031: 286-301.

Martin JE, Moskowitz M and Milbrath JR. Breast cancer missed by mammography. AJR 1979; 132: 737-739.

Metz CE and Shen S. Gains in diagnostic accuracy from replicated readings of diagnostic images: prediction and assessment in terms of ROC analysis. Med Decis Making 1992; 12: 60-75.

Metz CE, Shen JH, and Herman BA: New methods for estimating a binormal ROC curve from continuously distributed test results. Presented at the 1990 Annual Meeting of the American Statistical Association, Anaheim, CA.

Moskowitz M, *Benefit and risk. In: Breast Cancer Detection: Mammography and Other Methods in Breast Imaging*, 2nd ed. Eds. L. W. Bassett and R. H. Gold, (Grune and Stratton, New York, 1987).

Seidman H, Gelb SK, Silverberg E, LaVerda N and Lubera JA. Survival experience in the Breast Cancer Detection Demonstration Project. CA Cancer J Clin. 1987; 37: 258-290.

Shtern F, Stelling C, Goldberg B and Hawkins R. Novel technologies in breast imaging: National Cancer Institute perspective. Society of Breast Imaging, Orlando, Florida, 1995; 153-156.

Smart CR, Hendrick RE, Rutledge JH and Smith RA. Benefit of mammography screening in women ages 40 to 49 years: current evidence from randomized controlled trials. Cancer 1995; 75: 1619-1626.

Tabar L and Dean PB, *Teaching Atlas of Mammography*, (Thieme, New York, 1985).

Thurfjell EL, Lernevall KA and Taube AAS. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994; 191: 241-244.

Wallis MG, Walsh MT and Lee JR. A review of false negative mammography in a symptomatic population. Clinical Radiology 1991; 44: 13-15.

(11) Appendix

Publications enclosed

Journal Articles

1. Chan HP, Sahiner B, Helvie MA, Petrick N, Roubidoux MA, Wilson TE, Adler DD, Paramagul C, Newman JS, Sanjay-Gopal S. Improvement of radiologists' characterization of mammographic masses by computer-aided diagnosis: an ROC Study. Radiology 1999; 212: 817-827.

2. Petrick N, Chan HP, Sahiner B, Helvie MA, Goodsitt MM. Combined adaptive enhancement and object-based region growing for automated detection of masses on mammograms. Medical Physics 1999; 26: 1642-1654.

Conference Proceedings

1. Chan HP, Helvie MA, Petrick N, Sahiner B, Roubidoux MA, Wilson TE, Joynt LK, Hadjiiski LM, Paramagul C, Adler DD, Goodsitt MM. Digital Mammography: observer performance study of the effects of pixel size on radiologists' characterization of malignant and benign microcalcifications. Proc. SPIE 1999; 3659: 394-397.

2. Sahiner B, Chan HP, Petrick N, Wagner RF, Hadjiiski LM. Stepwise linear discriminant analysis in computer-aided diagnosis: the effect of finite sample size. Proc. SPIE 1999; 3661: 499-510.

3. Hadjiiski LM, Sahiner B, Chan HP, Petrick N, Helvie MA. Hybrid unsupervised-supervised approach for computerized classification of malignant and benign masses on mammograms. Proc. SPIE 1999; 3661: 464-473.

4. Dorfman DD, Berbaum KS, Lenth RV, Chen Y-F. Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: Split plot experimental design. Proc. SPIE 1999; 3663: 91-99.

Improvement of Radiologists' Characterization of Mammographic Masses by Using Computer-aided Diagnosis: An ROC Study[1]

Heang-Ping Chan, PhD
Berkman Sahiner, PhD
Mark A. Helvie, MD
Nicholas Petrick, PhD
Marilyn A. Roubidoux, MD
Todd E. Wilson, MD
Dorit D. Adler, MD
Chintana Paramagul, MD
Joel S. Newman, MD
Sethumadavan Sanjay-Gopal, PhD

Index terms:
Breast neoplasms, 00.31, 00.32
Breast neoplasms, radiography, 00.111, 00.119
Breast radiography, 00.111, 00.119
Computers, diagnostic aid
Receiver operating characteristic curve (ROC)

Radiology 1999; 212:817–827

Abbreviations:
CAD = computer-aided diagnosis
PPV = positive predictive value
ROC = receiver operating characteristic

[1] From the Department of Radiology, University of Michigan Hospital, UH B1F510, 1500 E Medical Center Dr, Ann Arbor, MI 48109-0030. From the 1997 RSNA scientific assembly. Received August 10, 1998; revision requested September 8; revision received November 30; accepted January 21, 1999. Supported in part by United States Public Health Service grant CA 48129 and by U.S. Army Medical Research and Materiel Command grant DAMD 17-96-1-6254. B.S. supported by Career Development award DAMD 17-96-1-6012 from the U.S. Army Medical Research and Materiel Command. N.P. supported by a grant from the Whitaker Foundation. **Address reprint requests to** H.P.C. (e-mail: *chanhp@umich.edu*).

The content of this article does not necessarily reflect the position of the funding agencies, and no official endorsement of any equipment or product of any companies mentioned in this article should be inferred.

© RSNA, 1999

Author contributions:
Guarantor of integrity of entire study, H.P.C.; study concepts and design, H.P.C., M.A.H., B.S., N.P.; literature research, H.P.C., M.A.H.; experimental studies, M.A.H., M.A.R., T.E.W., D.D.A., C.P., J.S.N.; data acquisition, all authors; data analysis, H.P.C., B.S., N.P.; statistical analysis, H.P.C.; manuscript preparation, editing, and review, H.P.C., B.S., M.A.H., N.P., M.A.R., T.E.W., D.D.A., C.P., J.S.N.

PURPOSE: To evaluate the effects of computer-aided diagnosis (CAD) on radiologists' classification of malignant and benign masses seen on mammograms.

MATERIALS AND METHODS: The authors previously developed an automated computer program for estimation of the relative malignancy rating of masses. In the present study, the authors conducted observer performance experiments with receiver operating characteristic (ROC) methodology to evaluate the effects of computer estimates on radiologists' confidence ratings. Six radiologists assessed biopsy-proved masses with and without CAD. Two experiments, one with a single view and the other with two views, were conducted. The classification accuracy was quantified by using the area under the ROC curve, A_z.

RESULTS: For the reading of 238 images, the A_z value for the computer classifier was 0.92. The radiologists' A_z values ranged from 0.79 to 0.92 without CAD and improved to 0.87–0.96 with CAD. For the reading of a subset of 76 paired views, the radiologists' A_z values ranged from 0.88 to 0.95 without CAD and improved to 0.93–0.97 with CAD. Improvements in the reading of the two sets of images were statistically significant (P = .022 and .007, respectively). An improved positive predictive value as a function of the false-negative fraction was predicted from the improved ROC curves.

CONCLUSION: CAD may be useful for assisting radiologists in classification of masses and thereby potentially help reduce unnecessary biopsies.

Breast cancer is the most prevalent non–skin cancer in women; 178,700 new cases are estimated to have occurred in 1998 (1). The mortality of breast cancer is the second highest among all cancer deaths in women (1). At present, there is no effective method to prevent breast cancer. The best approach to reducing the breast cancer mortality rate is early detection and treatment. Because the mammographic features of early-stage breast cancers are not very specific, the need for high detection sensitivity leads to biopsy of many low-suspicion lesions. The positive predictive values (PPVs) of mammographic signs are, therefore, often below 30% (2,3).

Computer-aided diagnosis (CAD) is considered to be one of the approaches that may improve the efficacy of mammography (4). With CAD, a computerized detection algorithm alerts a radiologist to the location of the suspicious lesions, and/or a trained computer classifier provides the radiologist with an estimate of the likelihood of malignancy of a lesion. The radiologist takes into consideration the information provided by the computer before making a decision. This "second opinion" may improve the diagnostic accuracy because it serves as a form of double reading (5). Furthermore, a computer evaluation is often more consistent and reproducible than a human decision maker (6).

Considerable research has been devoted to the development of computerized schemes for the detection and classification of mammographic abnormalities. These efforts have advanced the CAD technology such that clinical application appears to be possible in the

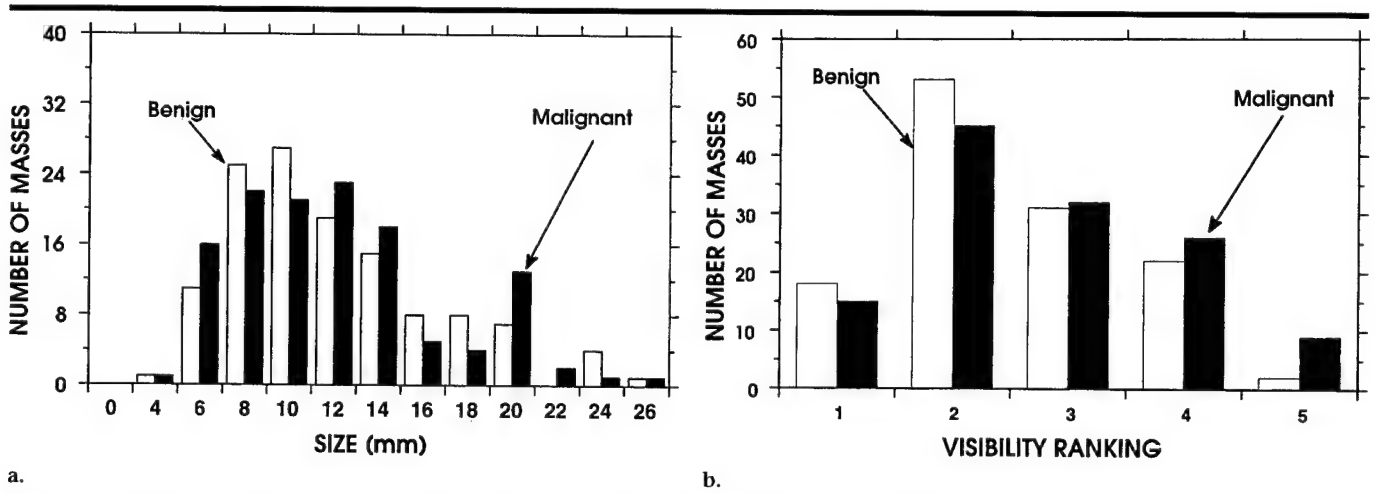


Figure 1. Histograms illustrate the distributions of **(a)** size (ie, length of the long axis) and **(b)** visibility ranking (1 = obvious, 5 = subtle) of the 253 masses included in the data set. Because classification accuracy depends on the case mix, these distributions provided some information on the masses in the data set.

near future. It is, therefore, necessary to evaluate the effects of CAD on radiologists' detection and diagnosis of mammographic lesions. In a previous receiver operating characteristic (ROC) study, we demonstrated that CAD could improve radiologists' accuracy in the detection of subtle microcalcifications on mammograms (7). Kegelmeyer et al (8) also reported an improvement in radiologists' sensitivity for the detection of spiculated masses with use of a computer aid. For the classification of mammographic lesions, it has been shown that a computer classifier that estimated the likelihood of malignancy on the basis of mammographic features extracted by radiologists could improve radiologists' accuracy in distinguishing malignant from benign lesions (9–11).

We previously conducted ROC studies to compare the performance of radiologists with that of the computer (12) and to compare radiologists' ability to classify masses with and without CAD (13). Jiang et al (14) also performed an ROC study of the effect of CAD on radiologists' performance in classifying microcalcifications. The results of all of these observer performance studies indicate the potential to improve mammographic interpretation with a computer aid.

We have developed an automated method to analyze masses seen on mammograms (15–17). A mass is segmented from its surrounding breast tissue, and an image transformation technique is used to transform the mass margin from the polar coordinate system to the Cartesian coordinate system. A linear discriminant classifier then extracts the useful texture features from the transformed image and merges them into a relative malignancy rating. Our approach is different from others that use a trained classifier to merge radiologist-extracted image features or feature codes by using the American College of Radiology Breast Imaging Reporting and Database System lexicon (9–11). Our fully automated method has the advantage that, unlike a human reader, it does not have variability in feature recognition and coding. In addition, the computer may be able to extract some information, such as texture features, that may not be readily perceived by human eyes. We conducted an ROC study to evaluate whether this computer aid can improve radiologists' performance in the classification of mammographic masses (13). The results of our observer performance study are described in this article.

Other investigators also have reported on automated algorithms for the classification of mammographic masses (18–21). The methods used in these algorithms varied, and their accuracy in classification cannot be compared directly because of the differences in the data sets. However, the effects of CAD on radiologists' performance are not expected to depend strongly on the specific algorithm if different computer aids of comparable accuracy are used. Therefore, the applications of the findings of this study should not be limited to our computerized classification aid.

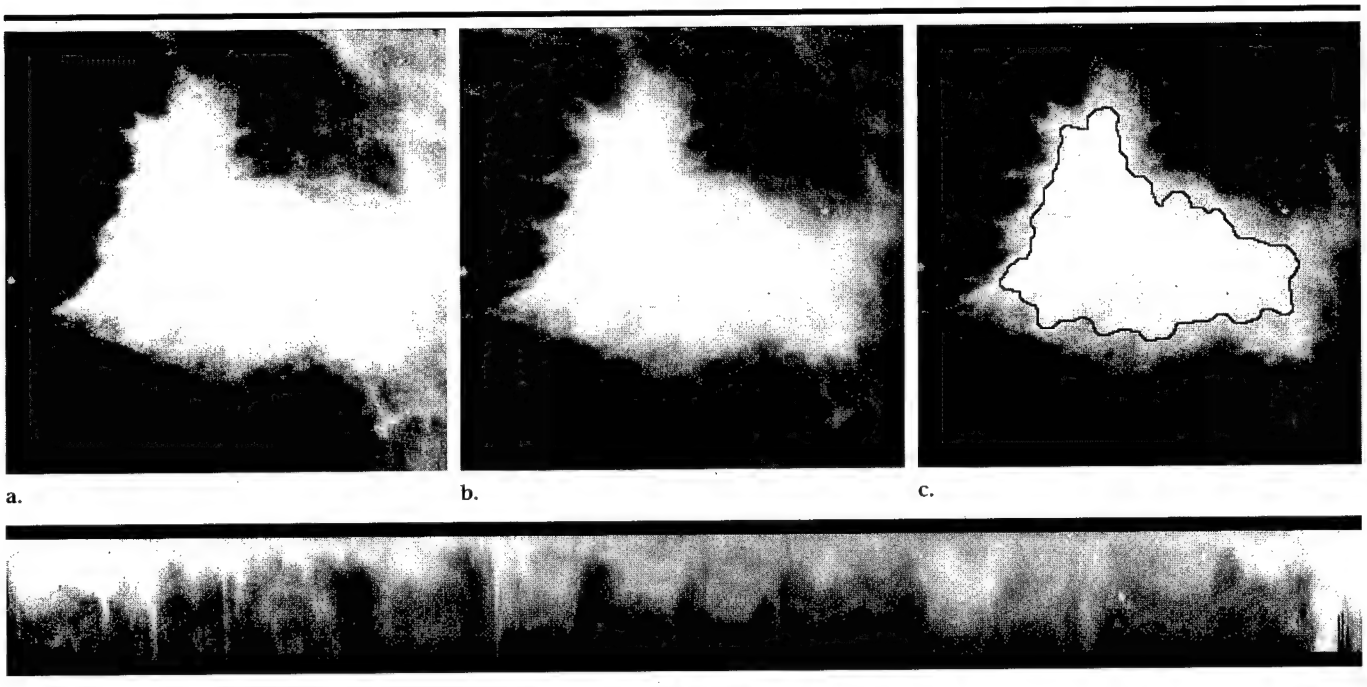
MATERIALS AND METHODS

Data Set

The data set for this study consisted of 253 mammograms obtained in 103 patients. Each image contained a biopsy-proved mass that was evaluated in this study. Some cases involved multiple views or images from multiple examinations. The cases were randomly selected from patient files from the breast imaging division of a National Cancer Institute–designated national cancer center with the approval of the Institutional Review Board. The PPV of masses recommended for biopsy at this center is about 25%–30%, but an approximately equal number of malignant and benign masses (127 and 126, respectively) were chosen to enhance the statistical power in this observer performance study. Any images that were judged to be technically poor were excluded.

The mammograms were acquired with a contact technique. The dedicated mammographic systems had a molybdenum anode and molybdenum filter, a 0.3-mm nominal focal spot, and a reciprocating grid. MinR/MinR-E screen-film systems (Eastman-Kodak, Rochester, NY) were used with these units. Sixty-two of the malignant masses and six of the benign masses were judged to be spiculated by a radiologist (M.A.H.) experienced in mammography. The radiologist also measured the size (ie, longest dimension) and ranked the visibility of the masses on a scale of 1 (obvious) to 5 (subtle) relative to the range of visibility of masses encountered in clinical practice. For a description of the masses included in the data set, histograms of the size and visibility of the masses are shown in Figures 1a and 1b, respectively.

For the computer analysis, the selected mammograms were digitized with a laser



a. b. c.

d.

Figure 2. Example of rubber-band-straightening transform for extraction of texture features in the margin region surrounding a mass. **(a)** Original and **(b)** background-corrected images showing the region of interest with the mass, **(c)** mammogram showing an outline of the segmented mass, and **(d)** rubber-band-straightening–transformed image of a 40-pixel-wide region surrounding the segmented mass.

imager (Lumisys DIS-1000, Los Altos, Calif) at a pixel size of 0.1 × 0.1 mm and 12-bit gray levels. This imager has an optical density range of about 0.0–3.5. The optical density on the film was digitized linearly to pixel value at a calibration of 0.001 optical density unit/pixel value in the optical density range of about 0.0–2.8. The digitizer deviated from a linear response at an optical density higher than 2.8.

For the observer experiments, we used laser-printed images of the digitized mammograms for all readings. The images were printed with a 969HQ laser imager (Imation, Oakdale, Minn) that was connected to a Macintosh computer (Apple Computer, Cupertino, Calif) through a special digital interface. The interface provided a 12-bit in, 10-bit out look-up table and allowed images to be scaled to different factors with 15 interpolation methods. Because this laser imager has a pixel size of about 0.085 mm, we enlarged the images by about 18% during printing to maintain them at the same size as the original mammograms. One of the interpolation methods was chosen by an experienced radiologist (M.A.H.), who inspected the printed images with a magnifier and evaluated the sharpness of the spicules and mass boundaries. Because of the small pixel size used for both digitization and printing, basically no noticeable blurring of the masses could be seen with the chosen interpolation method. The images were also inspected for the potential contouring effect of 10-bit output images, but no noticeable artifacts could be found. A linear pixel value–to–output optical density calibration curve of the laser imager was used for the printing. All images were printed with the same settings.

Computerized Classification of Masses

Our computerized method of classifying mammographic masses has been described in detail previously (15–17). The method is summarized as follows: A region of interest that contained the biopsy-proved mass was identified on the mammogram by the radiologist. Background correction based on a distance-weighted estimation method was applied to the region of interest to reduce the low-frequency density variation in the region. A median-filtered smoothed image and two high-frequency enhanced images were generated from the background-corrected region of interest. The smoothed and enhanced gray-level values at each pixel were used as features in a k-means clustering algorithm to classify the pixels into two clusters; one was the mass, and the other was the surrounding breast tissue background. By choosing an appropriate criterion, a mass region slightly smaller than the actual mass that was visible on the image was segmented.

The boundary of the segmented region was smoothed by morphologic filtering. A new image transformation technique, referred to as the rubber-band-straightening transform, was used to transform a 40-pixel-wide region that surrounded the segmented mass boundary into a rectangular region. After transformation, the mass margin became approximately parallel, and any spicules that were radiating from the mass became approximately perpendicular, to the long dimension of the rectangular region. The rubber-band-straightening transform enabled the spicules to be aligned approximately in a uniform direction and thus facilitated the extraction of texture features from the margin of the mass. An example of a rubber-band-straightening–transformed image is shown in Figure 2.

Two types of texture features were found to be useful for classification. The first set of features included eight texture measures derived from the spatial gray-level dependence matrices of the rubber-band-straightening–transformed image. A spatial gray-level dependence matrix ele-

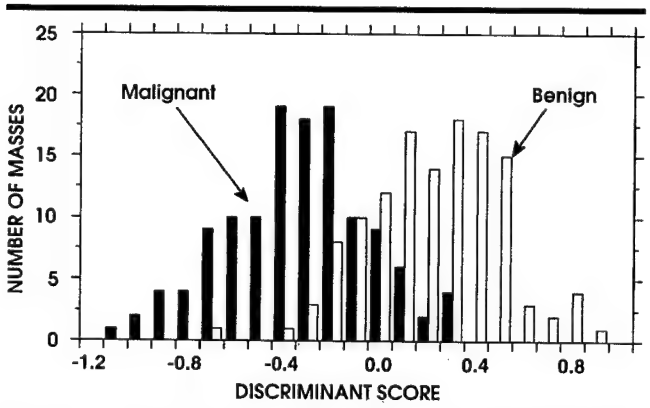


Figure 3. Histogram of the test discriminant scores of the 253 masses obtained from the linear discriminant classifier by using a "leave one case out" training and test resampling scheme. For this classifier, a smaller discriminant score corresponded to a higher likelihood of malignancy. The discriminant scores were used as the decision variable in the ROC analysis of classification performance.

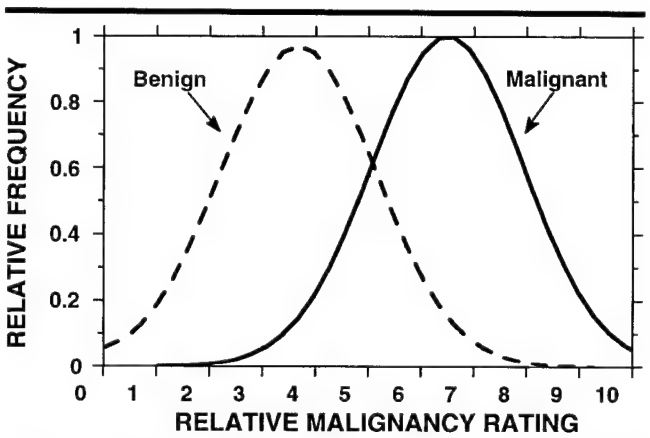


Figure 4. Binormal distribution fitted to the histogram of the discriminant scores of the malignant and benign masses. The discriminant scores were linearly transformed into a relative malignancy rating ranging from 1 to 10, where 1 corresponded to the most benign rating and 10 corresponded to the most malignant rating. This binormal distribution was shown to the observers during the training session to explain the rating scale of the computer classifier.

ment $p_{\theta,d}(i,j)$ is the joint probability of the occurrence of gray levels i and j for pixel pairs that are separated by a distance d and at a direction θ (22). For analysis of the masses, the spatial gray-level dependence matrices were constructed for 10 pixel distances (d = 1, 2, 3, 4, 6, 8, 10, 12, 16, 20 pixels) and in four directions (0°, 45°, 90°, 135°) relative to the mass boundary. Therefore, a total of 320 spatial gray-level dependence texture features were extracted.

The second set of texture features was derived from the run length statistics matrices of the horizontal and vertical gradient images of the rubber-band-straightening–transformed margin region. Five texture measures were extracted from the run length statistics matrix in each of the two directions (0° or 90°) on each gradient image. A total of 20 run length statistics texture features were thus obtained. Therefore, we had a total of 340 features from the two types of texture measures.

A stepwise linear discriminant feature selection procedure (23) was used to select the most effective features from the available feature set. A total of 41 features were selected. The selected features were input into the Fischer linear discriminant classifier (24) as predictor variables. A "leave one case out" resampling scheme was used to train and test the classifier. A histogram illustrating the test discriminant scores of the 253 masses is shown in Figure 3. For this classifier, a smaller discriminant score corresponded to a higher likelihood of malignancy. By using the test discriminant score as the decision variable, the performance of the computer classifier could be evaluated by using ROC analysis (17,25,26) and compared with that of the radiologists, as described later.

Relative Malignancy Rating of the Masses

For the observer performance study, we provided a relative malignancy rating of each mass to the observer during the reading session with CAD. The relative malignancy rating was obtained by taking a linear transformation of the computer classifier's decision variable to a range of 1–10 and rounding the value to the nearest integer. The transformation also reversed the relative magnitude of the decision variables so that 1 corresponded to the highest benignity rating, and 10 corresponded to the highest malignancy rating.

The purpose of the transformation was to provide a simple and intuitive relative scale for the observer. Because the transformation was linear and monotonic, the distributions of the normal and abnormal samples, as well as their ROC curves, were not affected, with the exception of a small error caused by making the decision variables discrete. Furthermore, the slope a and intercept b parameters that were fitted to the transformed discriminant scores for the normal and abnormal samples by using the LABROC program (26) were used to generate a binormal distribution. The fitted binormal distribution with the relative malignancy rating on a 1–10 scale (Fig 4), together with the computer's ROC curve, were shown and explained to the observers during a training session.

Observer Performance Study

Two ROC experiments (27) were conducted: The masses were evaluated from a single view in the first experiment and from two views in the second experiment. The location of the biopsy-proved mass was marked on each image so that the correct mass was evaluated by all observers. The observers were instructed to ignore any other possible masses on the images. Six radiologists (M.A.H., M.A.R., T.E.W., D.D.A., C.P., J.S.N.) who are approved by the Mammography Quality Standards Act and have 7–20 years of experience in interpreting mammograms participated in the observer performance experiments.

There were two reading sessions in each experiment—one with CAD and the other without CAD. The observers were asked to rate the likelihood of malignancy of the masses on a 10-point confidence rating scale under all reading conditions. In the first session, half the observers interpreted the images without CAD, and the other half interpreted them with CAD. The two reading sessions in the same experiment were separated by at least 3 weeks, and the two experiments were separated by 6 months. For all four reading sessions, the observer had unlimited time to read each case. To estimate the average reading time per case for each observer, the reading time for each case was recorded by using a stopwatch.

In the first experiment, the data set of 253 single-view mammograms was divided into a training set of 15 mammograms and a study set of 238 mammo-

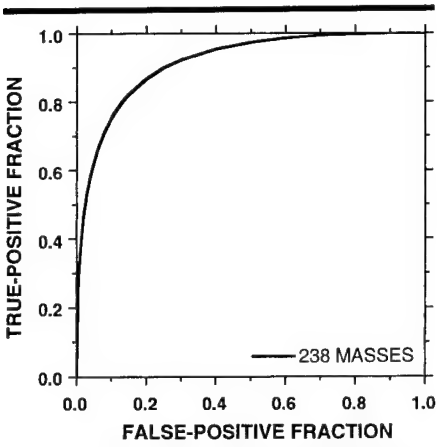


Figure 5. ROC curve for computerized classification of the 238 masses used in the observer performance study with single-view reading. The computer's ROC curve can be compared with the radiologists' ROC curves obtained from the single-view reading experiment illustrated in Figures 6 and 8.

grams (117 benign, 121 malignant). In each reading session, training was conducted before the reading of the study images. For the reading session with CAD, the fitted binormal distributions of the computer rating scores (Fig 4) for the entire data set were explained to the observer during training to familiarize the observer with the computer's rating scale. The computer rating of the mass was displayed on each image. After reading each training image, the observer was told the results of biopsy of the mass.

Each observer read the entire data set in one reading session. The order of the study images was randomized by a random number generator. The random sequence was different for each observer and for each reading session by the same observer. For the reading session with CAD, the observer was free to look at the computer rating, which was displayed on the image, either before or after estimating the likelihood of malignancy of the mass. However, each observer was asked to always read the computer rating before making a final decision. The observer was not informed of the pathologic results of any mass on the study images.

The second experiment was very similar to the first experiment. From the 238 single-view mammograms, 76 matched pairs (37 benign, 39 malignant) of craniocaudal and mediolateral oblique or lateral views were found. Another six pairs of two-view mammograms were identified from the rest of the images and used as training cases. The remaining mammograms were either single-view images or additional views of the pairs already chosen, so they were not used in this experiment. In this experiment, the observers were not informed of the pathologic results of any study case in any reading session. The 76 pairs of mammograms were read in one reading session by each observer.

For the reading session with CAD, the rating of the mass in each view was displayed on the respective image. The computer ratings of the mass on the two views were generally different. It was up to the observer to decide how to merge the two-view information. Observers were asked to give a single rating of the mass after reading both views.

ROC Analysis

The confidence ratings of each observer obtained from each reading condition were analyzed by using ROC methodology, and the classification accuracy was quantified by using the area under the ROC curve, A_z. A maximum likelihood estimation of the binormal distribution was fitted to the confidence ratings by using the LABROC program. This program provides an estimate of the A_z and of the a and b parameters of the ROC curve. The statistical significance of the difference in A_z between the reading with CAD and that without CAD was estimated with two methods: One was the Student paired t test for observer-specific paired data; the other was the Dorfman-Berbaum-Metz method for analysis of multireader, multicase ROC data (28). The statistical significance of the difference in $A_z$for reading single-view and two-view mammograms was estimated by using the Student paired t test for the six observers. The Student paired t test takes into account the statistical variation of readers, whereas the Dorfman-Berbaum-Metz method considers both reader variation and case sample variation by means of an analysis of variance approach. Therefore, the results of Dorfman-Berbaum-Metz analysis can be generalized to the population of readers as well as to the population of case samples.

Positive Predictive Value

An ROC curve represents the entire range of operating conditions of a diagnostic process and is independent of disease prevalence. When the disease prevalence is known, any operating point on an ROC curve can be used to derive the PPV and the corresponding false-negative fraction (false-negative fraction = 1 − true-positive fraction) on the basis of the following relationship: PPV = TPF × P(M)/[TPF × P(M) + FPF × P(B)], where TPF is the true-positive fraction, FPF is the false-positive fraction at the chosen decision threshold, and P(M) and P(B) are the prevalences of malignant and benign cases, respectively. By varying the decision threshold, the dependence of the PPV on the false-negative fraction can be derived.

Because our data set did not include masses on which biopsy had not been performed, the ROC curves obtained in this study cannot be generalized to predict the performance of the computer classifier and radiologists in clinical practice. However, to demonstrate the possible effect of CAD on the PPV in the population of masses in which biopsy is likely to be performed under the current clinical criteria, we can estimate the PPV by using the prevalence of the malignant and benign masses in this patient group. Because the PPV of masses sent for biopsy ranges from about 25% to 44% in general and from about 25% to 30% at our institution, for the purposes of our estimation, we assumed that the P(M) was 25% and the P(B) was 75% in this population. A higher prevalence of malignant cases would cause an increase in the PPV, but the trend between the PPV curves with and without CAD would be similar.

RESULTS

The ROC curve illustrating the performance of the computer classifier for the 238 study mammograms is shown in Figure 5. The ROC curve for the entire set of 253 mammograms (not shown) was almost identical to that of the 238 study cases; this indicates that the 15 training cases were typical of the 238 cases used in the study. The A_z values (± SD) for both ROC curves were 0.92 ± 0.02.

For the first experiment of reading the 238 single-view mammograms, the ROC curves for the readings by the six radiologists both without and with CAD are shown in Figures 6a and 6b, respectively. The A_z values of the six radiologists for the readings with and without CAD are listed in Table 1.

For the second experiment of reading the 76 pairs of two-view mammograms, the ROC curves for the readings by the six radiologists both without and with CAD are shown in Figures 7a and Figure 7b, respectively. The A_z values of the six radiologists in this experiment are also listed in Table 1.

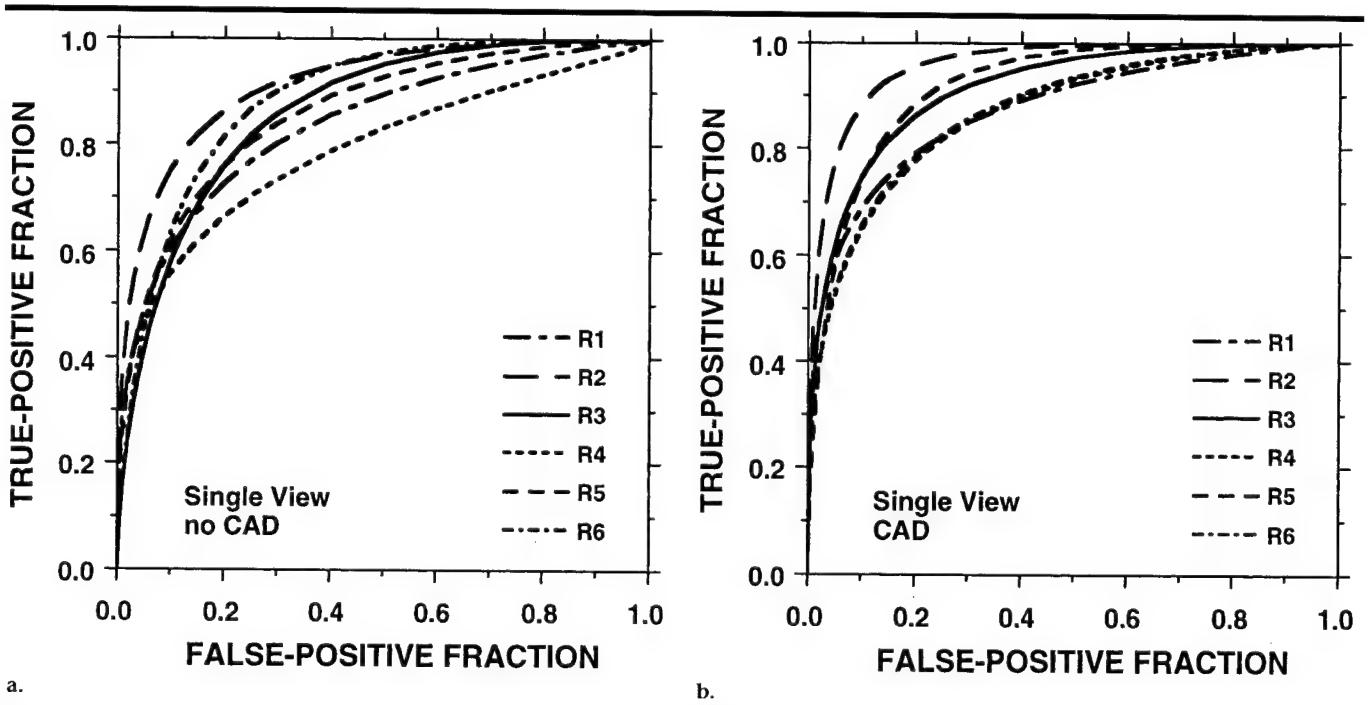


Figure 6. ROC curves for the six observers for single-view reading of the masses (**a**) without CAD and (**b**) with CAD. (**a, b**) *R1* = reader 1, *R2* = reader 2, *R3* = reader 3, *R4* = reader 4, *R5* = reader 5, *R6* = reader 6. Five of the six observers achieved an increase in the area under the ROC curve, A_z, with CAD.

The average ROC curve was derived from the average *a* and *b* parameters of the six individual ROC curves for a given reading condition (27). The average ROC curves for the four reading conditions are shown in Figure 8. The A_z values of the average ROC curves are listed in Table 1.

For the reading of the single-view mammograms, the performance of the computer classifier was comparable to that of the radiologist (reader 2) who had the highest classification accuracy (compare Figs 5 and 6) and higher than the average performance of the six radiologists (compare Figs 5 and 8). When the radiologists read the images with the computer aid, the classification accuracy of five radiologists improved (Table 1); the improvement in their A_z values ranged from 0.04 to 0.08. The average performance of the six radiologists became comparable to that of the computer classifier. The improvement in the radiologists' classification accuracy by using CAD was statistically significant ($P = .022$, Student paired *t* test; $P = .020$, Dorfman-Berbaum-Metz method). Reader 2 with CAD obtained an A_z value of 0.96, which was higher than that obtained by the radiologist alone or by the computer alone.

A trend similar to that with the single-view readings was observed with the two-view readings. The A_z value of the computer classifier for the corresponding 152 single-view masses was 0.91 ± 0.02. The classification accuracy of all six radiologists improved when they read the mammograms with the computer aid. The increase in the A_z values ranged from 0.01 to 0.07. The improvement was statistically significant ($P = .007$, Student paired *t* test; $P = .026$, Dorfman-Berbaum-Metz method). With CAD, two radiologists achieved an A_z value of 0.97, which was higher than that obtained by the radiologists alone or by the computer alone. These results indicate that the second opinion provided by the computer classifier might have strengthened the radiologists' confidence in the interpretation of some difficult cases but had less influence on the radiologists' decision when the computer made mistakes or when the radiologists were confident about their decision.

As can be seen from the data in Table 1,

TABLE 1
Areas under the ROC Curves for the Classification of Masses with and without CAD by the Six Radiologists

Radiologist No.	A_z (Single View)*		A_z (Two View)†	
	Without CAD	With CAD	Without CAD	With CAD
1	0.84 ± 0.03	0.87 ± 0.02	0.90 ± 0.03	0.93 ± 0.03
2	0.92 ± 0.02	0.96 ± 0.01	0.95 ± 0.02	0.97 ± 0.02
3	0.86 ± 0.02	0.91 ± 0.02	0.92 ± 0.03	0.93 ± 0.03
4	0.79 ± 0.03	0.87 ± 0.02	0.88 ± 0.04	0.95 ± 0.03
5	0.86 ± 0.02	0.92 ± 0.02	0.93 ± 0.03	0.97 ± 0.02
6	0.89 ± 0.02	0.87 ± 0.02	0.89 ± 0.04	0.93 ± 0.03
A_z from average a, b parameters	0.87	0.91	0.92	0.96

Note.—Data are the mean ± SD.

* $P = .022$ for the difference between the A_z values measured with CAD and those measured without CAD, as determined by using the Student two-tailed *t* test. $P = .020$ for this difference, as determined by using the Dorfman-Berbaum-Metz method.

† $P = .007$ for the difference between A_z values measured with CAD and those measured without CAD, as determined by using the Student two-tailed *t* test. $P = .026$ for this difference, as determined by using the Dorfman-Berbaum-Metz method.

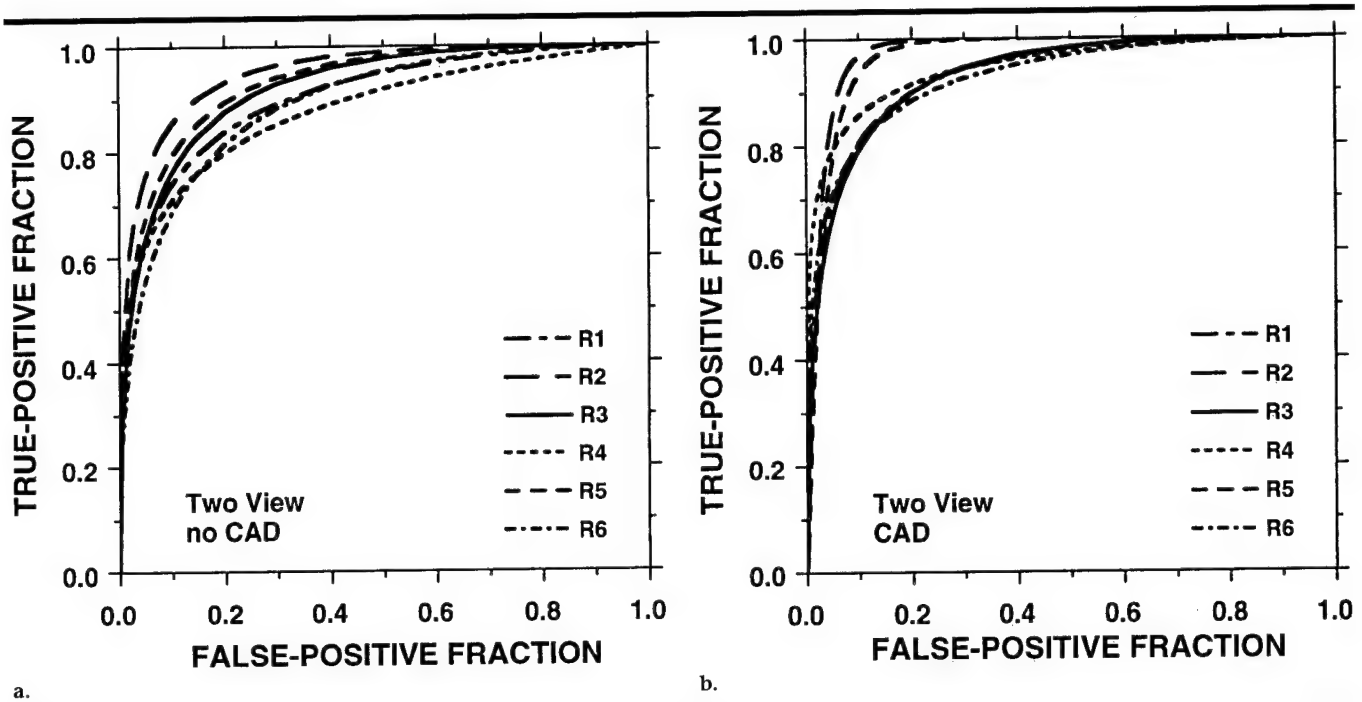


Figure 7. ROC curves for the six observers for two-view reading of the masses (**a**) without CAD and (**b**) with CAD. (**a, b**) *R1* = reader 1, *R2* = reader 2, *R3* = reader 3, *R4* = reader 4, *R5* = reader 5, *R6* = reader 6. All six observers achieved an increase in the area under the ROC curve, A_z, with CAD.

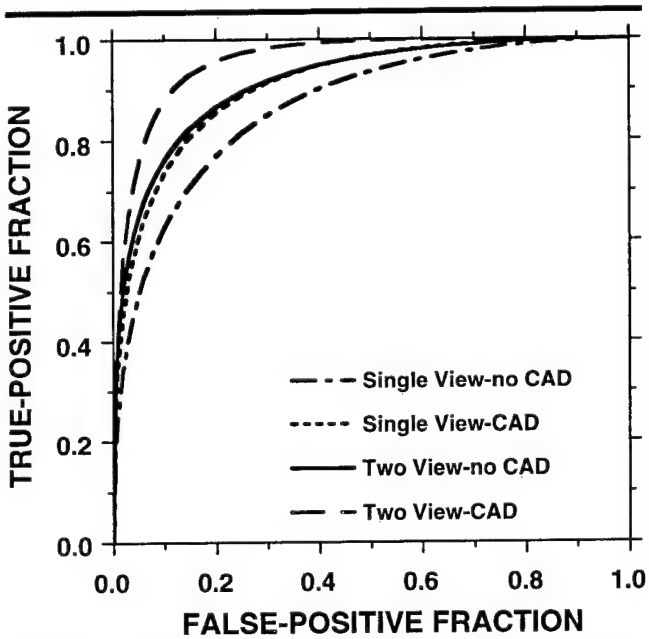


Figure 8. Average ROC curve obtained from the average *a* and *b* parameters of the six individual ROC curves for each of the four reading conditions. An improved ROC curve was achieved with CAD in both the single-view and two-view reading experiments.

the radiologists' accuracy in classifying masses by reading two-view mammograms was consistently higher than that by reading single-view mammograms ($P = .008$). This trend remained when they read the mammograms with CAD ($P = .007$). These findings are consistent with the clinical experience of the radiologists that at least two views of mammograms are needed to effectively evaluate a suspicious lesion.

The PPV as a function of the false-negative fraction was derived from the fitted ROC curves under the assumption that the prevalence of malignant masses was 25% in the population of masses sent for biopsy. The PPVs estimated for the six observers who read the two-view mammograms with and without CAD are plotted in Figure 9. CAD would provide an improvement in the PPV in the high false-negative fraction range for all observers except readers 2 and 5. The increase in the PPV at a decision threshold of "no missed malignant mass" (ie, false-negative fraction = 0) varied over a wide range; the largest gain, 39%, would be achieved by reader 2, and the smallest gain, 0%, would be achieved by reader 4.

DISCUSSION

In the observer experiment of reading two-view mammograms with CAD, we presented the computer's rating of each view separately. The decision of how to merge the computer ratings of the two views was left to the radiologist. It is likely that the radiologists took the conservative approach of using the highest malignancy rating of the two as the computer's overall rating. However, it also might have depended on whether the relative ranking between the two computer ratings agreed with the observer's opinion. In some cases, we observed that the radiologist's rating was very different from the computer's rating of either view.

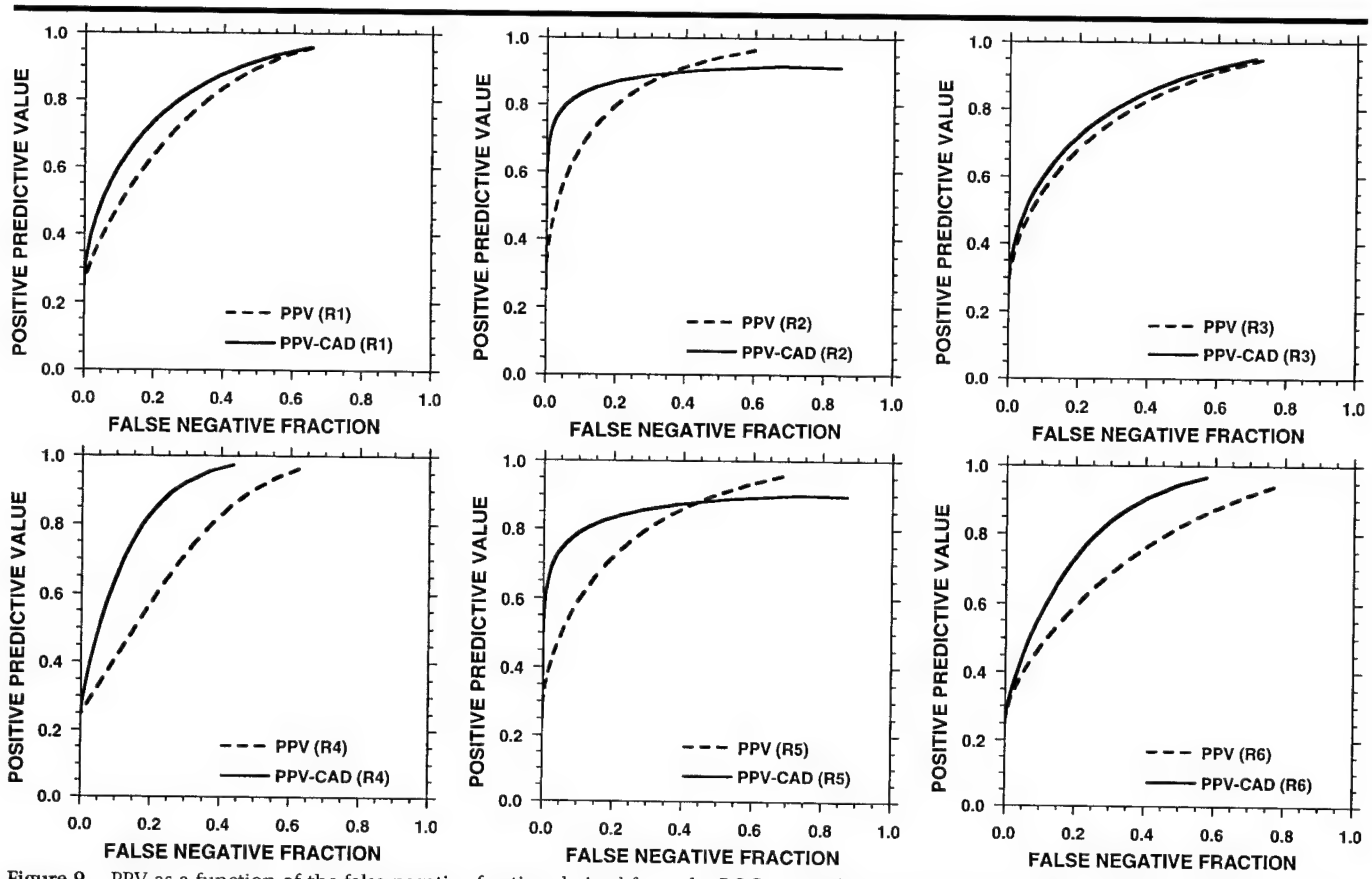


Figure 9. PPV as a function of the false-negative fraction derived from the ROC curves for the six observers (Fig 7). The PPV was predicted for a population of masses in which biopsy was likely to be performed under current clinical criteria and by assuming the prevalence of malignant masses to be 25%. *R1* = reader 1, *R2* = reader 2, *R3* = reader 3, *R4* = reader 4, *R5* = reader 5, *R6* = reader 6.

Because decision making is a complex process, the simple approach of using the highest malignant rating or the average rating from multiple views may not be the method preferred by radiologists. The separate ratings that we used in this study would provide less biased information. Further investigation is needed to determine the best approach of presenting the computer's ratings to radiologists in clinical practice.

To obtain insight into how the radiologists might use the two-view information, we compared the classification results from their true two-view reading with those from a simulated two-view reading without the computer aid. The latter results were derived from ratings of single-view readings of the same 76 pairs of mammograms interpreted in experiment 2 by assuming two strategies—one in which the highest malignancy rating between the two ratings was used, and the other in which the average of the two ratings was used (Table 2). The A_z values for these classification ratings derived from the single-view reading are listed in Table 2. The corresponding A_z values for the computer classifier are also given in Table 2 for comparison.

The A_z values for the maximal rating and the average rating were similar. Four of the radiologists obtained higher A_z values at the true two-view reading; the A_z values obtained by the remaining two radiologists were lower than those obtained at the simulated two-view reading. Although the difference did not achieve statistical significance ($P = .37$) and both readings included intraobserver variations, there seemed to be a slight trend toward the true two-view reading being more accurate than the simulated two-view reading. This may indicate that the radiologists used a more complex decision-making process to interpret the two views of the masses than that of simply maximizing or averaging the ratings from each view.

In this study, the discriminant scores of the masses given by the computer classifier were transformed into a relative malignancy rating. The relative malignancy rating scale and the distribution of the malignant and benign masses along the relative rating scale were explained to the observers in the training sessions. A relative malignancy rating scale was used because the true likelihood of malignancy of the masses could not be estimated from a small data set, as will be explained. However, the relative rating scale provided by the computer was ad-

TABLE 2
Estimation of the Malignancy Classification of 76 Masses by Two-View Reading, as Simulated from Single-View Reading of Mammograms by Radiologists without CAD

Radiologist No.	A_z	
	Maximal Rating	Average Rating
1	0.94 ± 0.03	0.93 ± 0.03
2	0.94 ± 0.03	0.94 ± 0.03
3	0.84 ± 0.05	0.86 ± 0.04
4	0.85 ± 0.04	0.83 ± 0.05
5	0.88 ± 0.04	0.89 ± 0.04
6	0.91 ± 0.03	0.92 ± 0.03
Computer	0.96 ± 0.02	0.96 ± 0.02

Note.—Data are the mean ± SD. Two strategies were used: In one, the highest of the malignancy ratings on each view was used; in the other, the average between the two ratings was used.

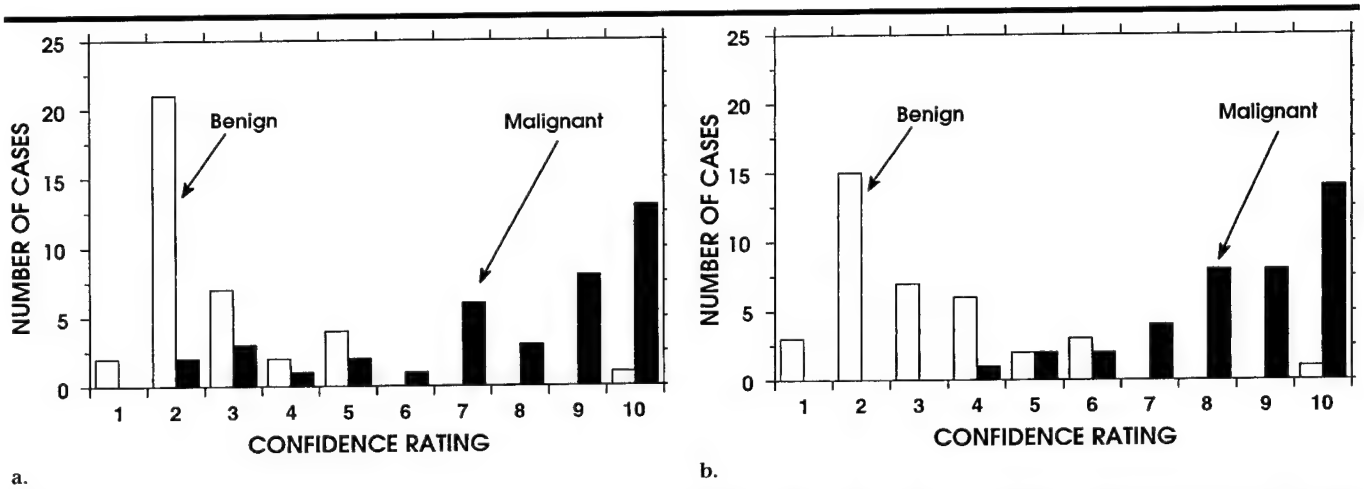


Figure 10. Histograms illustrate the confidence ratings of reader 5 obtained by reading 76 two-view mammograms (a) without CAD and (b) with CAD. The specificity of reader 5 at 100% sensitivity would increase from 5% (two of 37 masses) without CAD to 68% (25 of 37 masses) with CAD if an appropriate decision threshold were chosen.

equate for measuring the relative performance of classification with and without CAD in an ROC study.

If a computer classifier is trained and tested with very large data sets, and if both the malignant and benign cases represent random samples of the population, then the likelihood of malignancy of a classified mass can be estimated on the basis of the probability distributions of the classifier's test output scores and the prevalence of the two classes of masses in the patient population. However, with a relatively small data set, such as that used in this and other observer studies (14), there are limitations. First, the performance of a classifier trained with a small sample set may have large bias and variance (29–31). Second, the data set in this study did not include masses on which biopsy was not performed, so it did not represent a random sample of the masses in the patient population. If our classifier were applied to all cases of solid masses in clinical practice, the probability distribution of the test scores for the two classes of masses would be different from that of the current data set.

If we ignore the patient population at large, it is possible to estimate the likelihood of malignancy of a mass on the basis of the probability distribution of the classifier output scores by using the prevalence of the two classes of masses in this specific data set. However, the likelihood of malignancy derived in this way will be completely different from the true likelihood of malignancy of a mass in the patient population. This can be easily seen if one considers that the same mass with the same discriminant score will have a smaller likelihood of malignancy if it is analyzed within a data set that has a lower prevalence of malignant cases than that in the current data set.

Training the participating radiologists with a "likelihood of malignancy" derived from a small data set for the observer experiment may mislead them if they encounter a similar mass in their clinical practice. We, therefore, preferred to use a "relative malignancy rating," which is independent of the prevalences of malignant and benign masses in the data set. As long as the same classifier and the same linear transformation are used for classifying masses, the relative malignancy rating for a given mass will remain the same, regardless of the types of other masses in the data set. When a computer classifier is implemented in a clinical setting and its performance can be established in the patient population, the true likelihood of malignancy of a given mass can be estimated and provided to the radiologist. The true likelihood of malignancy may be a more informative measure for radiologists in the clinical application of CAD.

For the reading of the 76 two-view mammograms, the results of the ROC study indicated an improvement in the A_z value for all six radiologists when the computer aid was used. This indicates an overall increase in the separation of confidence rating distributions between the malignant and benign cases. The histograms in Figure 10 illustrate the distributions of confidence ratings with and without CAD for reader 5, who achieved the second greatest improvement in both the A_z value (Table 1) and the separation of malignant from benign distributions. Without CAD, this reader's ratings of the malignant cases ranged from 2 to 10. This is consistent with the fact that biopsy was performed in all masses in the data set to avoid missing the malignant cases. With CAD, there was marked improvement in the separation of the two distributions. It is possible to set a decision threshold at a confidence rating of 4, below which biopsy would not need to be performed and no malignant masses would be missed. The number of benign masses that could be identified without missing a malignant mass by setting an appropriate threshold would increase by 23 (out of 76 cases) for reader 5. Five of the six radiologists in our ROC study achieved an improvement in distinguishing benign from malignant masses, and one radiologist had no difference. Although the improvement of the five radiologists varied over a wide range, from one to 25 cases, this result indicates a strong possibility that CAD can be used to reduce the number of unnecessary biopsies.

The large variation in improvement among the radiologists may have been due to the different degrees of confidence that they had in the computer aid. As with any new diagnostic tool, this confidence is influenced by the experience the radiologist has with the tool. Although the radiologists received training before the reading sessions, the high variability in confidence was not unexpected, because this ROC study was the first instance in which they had worked with the computer aid. Their confidence levels may have also been reflected in the relatively low accuracy of classification by some radiologists with CAD compared with that of the computer classifier alone.

If a radiologist can increase his or her

confidence in the performance of a computer aid by gaining more extensive clinical experience, then he or she will likely be able to find the most effective way of merging his or her judgment with the computer's rating and thus reduce both interobserver and intraobserver variability. Because a radiologist who uses CAD can establish a meaningful decision threshold for biopsy only after becoming familiar with the sensitivity and specificity of working with CAD, the radiologists in this study were not asked to decide whether biopsy should have been performed on a mass. Rather, we focused on the evaluation of changes in the sensitivity and specificity of the radiologists' classification of masses when CAD was used.

In this ROC study, all six observers were attending radiologists with extensive experience in the interpretation of mammograms. It is possible that the computer aid may be even more useful to radiology residents or radiologists with less experience in mammography. The effect of CAD on mammographic interpretation by less-experienced readers will be a subject of investigation in future studies.

The observers were allowed unlimited time to read each case in this ROC study. To obtain an estimate of the change in reading time with CAD, we recorded the reading time of each observer in each reading session by using a stopwatch. For the single-view reading experiment, the average reading time per image without CAD varied from 4.3 seconds to 17.1 seconds (mean time for the six observers, 7.8 seconds). The average reading time per image with CAD varied from 4.2 seconds to 17.3 seconds (mean time, 7.3 seconds). For the two-view reading experiment, the average reading time per pair of images without CAD varied from 6.6 seconds to 16.0 seconds (mean time, 10.4 seconds). The average reading time per pair of images with CAD varied from 7.6 seconds to 27.1 seconds (mean time, 13.5 seconds).

The reading time essentially did not change with use of the computer aid for the single-view readings. For the two-view readings, the radiologists took longer with CAD, probably because they had to merge the two computer ratings and merge the computer ratings with their own evaluations. Further investigation is needed to determine whether there is a trade-off between the radiologist's efficiency and the method of presenting the computer rating and whether the reading time with CAD will depend on the experience that the radiologist has with the computer information.

In the observer study, we used laser-printed mammograms instead of the original mammograms for the reading experiments. A major reason is that it is difficult to keep all the original mammograms together for the entire period of the study because they are part of active patient files and thus often recalled for comparison with new studies or for other clinical reasons. Because the maximum optical density of laser-printed images was 3.1 for the laser imager used, the contrast on the printed mammograms was about 20% lower than that on the original mammograms. Although the image quality was slightly lower than that of the original, the laser-printed digitized images were judged to be adequate for reading the details of the masses by the participating radiologists. The laser-printed image set might also be considered as one that had slightly more subtle masses than the original set of images. Because the relative performance of two modalities is measured in ROC experiments, and because the readings both with and without CAD in this study were conducted with the same set of printed images, the relative performance of the two readings should be valid. It should also be noted that in order for a computer aid that uses automated image analysis to be widely accepted, direct digital mammography would have to be the imaging modality in clinical use. Laser-printed images or soft-copy monitors will be the display medium for the digital mammograms. The use of laser-printed images for this ROC study was therefore practical.

In our observer performance experiment, we found that CAD improved the radiologists' ability to distinguish malignant and benign masses. This is consistent with the results of other studies (11,14) in which a statistically significant improvement ($P < .001$ in both studies) in the radiologists' classification accuracy by using CAD was found. The results of the former study (11) further showed that the PPV of a recommendation for biopsy by the radiologists was significantly increased ($P < .001$). In our approach, the computer classifier automatically extracted image features, whereas in the other studies, the computer classifier used the radiologist's evaluation and other patient information as input. Therefore, it appears that CAD can provide a useful second opinion to radiologists, either by consistently extracting and analyzing the image features or by optimally weighting various diagnostic factors and thereby improving the consistency in the decision-making process. This suggests that a computer classifier that combines both approaches—that is, automatically extracts image features and optimally merges them with the radiologist's evaluation and patient information—may be even more effective for breast cancer diagnosis. The latter step will also improve the radiologist's utilization of the computer rating on the basis of the computer-extracted features; this utilization was found to have large interobserver variation in our ROC experiment.

In conclusion, an ROC study of the effects of CAD on radiologists' classification of malignant and benign masses on mammograms was conducted. The results showed that CAD can provide a statistically significant improvement in the classification accuracy—that is, in the A_z value—for both single-view reading ($P = .022$) and two-view reading ($P = .007$). The improved separation between the confidence ratings of the malignant masses and those of the benign masses indicates the potential that CAD may reduce the rate of biopsy of benign masses when decision thresholds are properly chosen by the radiologists. The decision threshold may vary among radiologists, as in the case of mammographic interpretation without CAD, and can be set after the radiologist working with CAD has established his or her sensitivity and specificity with this approach through clinical experience.

Further studies are needed to evaluate the effects of CAD on the accuracy of radiologist classification of masses in clinical settings in which the prevalence of malignant masses is different from that in a laboratory data set and the likelihood of malignancy of a mass can be estimated by the computer classifier. In the two-view reading ROC experiment, the reading time per case increased by about 30% with the use of CAD. The dependence of the radiologist's efficiency in reading with CAD on the presentation method and on the reader's experience in using the computer information also warrants further investigation.

Acknowledgments: The authors are grateful to Charles E. Metz, PhD for useful discussions and for the use of the LABROC and LABMRMC programs.

References

1. Landis SH, Murray T, Bolden S, Wingo PA. Cancer statistics 1998. CA Cancer J Clin 1998; 48:6–29.
2. Adler DD, Helvie MA. Mammographic biopsy recommendations. Curr Opin Radiol 1992; 4:123–129.

3. Kopans DB. The positive predictive value of mammography. AJR 1991; 158:521–526.
4. Shtern F. Digital mammography and related technologies: a perspective from the National Cancer Institute. Radiology 1992; 183: 629–630.
5. Thurfjell EL, Lernevall KA, Taube AAS. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994; 191:241–244.
6. Vyborny CJ. Can computers help radiologists read mammograms? Radiology 1994; 191:315–317.
7. Chan HP, Doi K, Vyborny CJ, et al. Improvement in radiologists' detection of clustered microcalcifications on mammograms: the potential of computer-aided diagnosis. Invest Radiol 1990; 25:1102–1110.
8. Kegelmeyer WP, Pruneda JM, Bourland PD, Hillis A, Riggs MW, Nipper ML. Computer-aided mammographic screening for spiculated lesions. Radiology 1994; 191: 331–337.
9. Getty DJ, Pickett RM, D'Orsi CJ, Swets JA. Enhanced interpretation of diagnostic images. Invest Radiol 1988; 23:240–252.
10. D'Orsi CJ, Getty DJ, Swets JA, Pickett RM, Seltzer SE, McNeil BJ. Reading and decision aids for improved accuracy and standardization of mammographic diagnosis. Radiology 1992; 184:619–622.
11. Baker JA, Kornguth PJ, Lo JY, Floyd CE. Artificial neural network: improving the quality of breast biopsy recommendations. Radiology 1996; 198:131–135.
12. Chan HP, Sahiner B, Petrick N, et al. Observer performance study of radiologists' reading of mammographic masses and comparison with computerized classification (abstr). Radiology 1996; 201(P):370.
13. Chan HP, Sahiner B, Helvie MA, et al. Effects of computer-aided diagnosis (CAD) on radiologists' classification of malignant and benign masses on mammograms: an ROC study (abstr). Radiology 1997; 205(P):275.
14. Jiang Y, Nishikawa R, Schmidt RA, Metz CE, Doi K. Improving breast cancer diagnosis with computer-aided diagnosis (CAD): an observer study (abstr). Radiology 1997; 205(P):274.
15. Sahiner B, Chan HP, Petrick N, Helvie MA, Adler DD, Goodsitt MM. Classification of masses on mammograms using rubberband straightening transform and feature analysis. Proc SPIE 1996; 2710:44–50.
16. Sahiner B, Chan HP, Petrick N, Helvie MA, Goodsitt MM. Computerized characterization of masses on mammograms: the rubber-band straightening transform and texture analysis. Med Phys 1998; 25:516–526.
17. Sahiner B, Chan HP, Petrick N, Helvie MA, Goodsitt MM. Design of a high-sensitivity classifier based on a genetic algorithm: application to computer-aided diagnosis. Phys Med Biol 1998; 43:2853–2871.
18. Ackerman LV, Gose EE. Breast lesion classification by computer and xeroradiograph. Cancer 1972; 30:1025–1035.
19. Kilday J, Palmieri F, Fox MD. Classifying mammographic lesions using computerized image analysis. IEEE Trans Med Imaging 1993; 12:664–669.
20. Pohlman S, Powell KA, Obuchowshi NA, Chilote WA, Grundfest-Broniatowski S. Quantitative classification of breast tumors in digitized mammograms. Med Phys 1996; 23:1337–1345.
21. Huo Z, Giger ML, Vyborny CJ, Wolverton DE, Schmidt RA, Doi K. Automated computerized classification of malignant and benign masses on digitized mammograms. Acad Radiol 1998; 5:155–168.
22. Haralick RM, Shanmugam K, Dinstein I. Texture features for image classification. IEEE Trans Syst Man Cybernetics 1973; 3:610–621.
23. Norusis MJ. SPSS for Windows release 6: professional statistics. Chicago, Ill: Statistical Product for Service Solutions, 1993.
24. Lachenbruch PA. Discriminant analysis. New York, NY: Hafner, 1975; 8–19.
25. Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986; 21:720–733.
26. Metz CE, Herman BA, Shen JH. Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data. Stat Med 1998; 17:1033–1053.
27. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 1989; 24:234–245.
28. Dorfman DD, Berbaum KS, Metz CE. ROC rating analysis: generalization to the population of readers and cases with the jackknife method. Invest Radiol 1992; 27:723–731.
29. Fukunaga K, Hayes RR. Effects of sample size on classifier design. IEEE Trans Pattern Analysis and Machine Intelligence 1989; 11:873–885.
30. Chan HP, Sahiner B, Wagner RF, Petrick N, Mossoba J. Effects of sample size on classifier design: quadratic and neural network classifiers. Proc SPIE 1997; 3034:1102–1113.
31. Chan HP, Sahiner B, Wagner RF, Petrick N. Classifier design for computer-aided diagnosis in mammography: effects of finite sample size. Med Phys 1997; 24:1034–1035.

Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms

Nicholas Petrick, Heang-Ping Chan, Berkman Sahiner, and Mark A. Helvie
The University of Michigan, Department of Radiology, CGC B2102, 1500 East Medical Center Drive, Ann Arbor, Michigan 48109-0904

(Received 15 July 1998; accepted for publication 27 April 1999)

As an ongoing effort to develop a computer aid for detection of masses on mammograms, we recently designed an object-based region-growing technique to improve mass segmentation. This segmentation method utilizes the density-weighted contrast enhancement (DWCE) filter as a preprocessing step. The DWCE filter adaptively enhances the contrast between the breast structures and the background. Object-based region growing was then applied to each of the identified structures. The region-growing technique uses gray-scale and gradient information to adjust the initial object borders and to reduce merging between adjacent or overlapping structures. Each object is then classified as a breast mass or normal tissue based on extracted morphological and texture features. In this study we evaluated the sensitivity of this combined segmentation scheme and its ability to reduce false positive (FP) detections on a data set of 253 digitized mammograms, each of which contained a biopsy-proven breast mass. It was found that the segmentation scheme detected 98% of the 253 biopsy-proven breast masses in our data set. After final FP reduction, the detection resulted in 4.2 FP per image at a 90% true positive (TP) fraction and 2.0 FPs per image at an 80% TP fraction. The combined DWCE and object-based region growing technique increased the initial detection sensitivity, reduced merging between neighboring structures, and reduced the number of FP detections in our automated breast mass detection scheme. © *1999 American Association of Physicists in Medicine.* [S0094-2405(99)00808-1]

Key words: computer-aided diagnosis, digital mammography, breast mass detection, density-weight contrast enhancement, region growing

I. INTRODUCTION

Mammographic screening has proven to be an effective method for early detection of breast cancer. Women in a regular mammographic screening program have a statistically significant reduction in breast cancer mortality when compared to women not in such a program.[1] In addition, independent double reading by two radiologists has proven to significantly increase the sensitivity of mammographic screening.[2] Therefore, regular screening and double reading would appear to be a sensible approach for breast cancer detection. While regular screening is emphasized in health care programs, the higher cost and increased workload on the radiologists may make double reading by two radiologists impractical in a general screening situation. Computer-aided diagnosis (CAD) is one alternative that could allow a large number of mammograms to be double read by a single radiologist aided by the computer. This technique may improve the accuracy of both detection and characterization of breast lesions.

Many researchers have been interested in computerized analysis of mammograms[3] and a number of groups have developed algorithms for automated detection of breast masses. The detection of spiculated masses has been of particular importance because of its high likelihood of malignancy. Karssemeijer *et al.*,[4] Kobatake *et al.*,[5] and Kegelmeyer *et al.*[6] have all proposed methods for detecting spiculated masses on digitized mammograms. However, since a number of malignant masses are not spiculated, other groups have tackled the general problem of identifying all types of breast masses on digitized mammograms.[3,7–11]

Our research group has reported on a method for automatically detecting masses on digitized mammograms.[10,12] The method employed multiple stages of density-weighted contrast enhancement (DWCE) segmentation. The DWCE segmentation was first applied to the full mammogram, and then reapplied to local regions within the mammogram to improve object border definition. A final object splitting stage was employed to eliminate merging between neighboring or overlapping breast structures. False positive (FP) reduction based on extracted morphological features was applied after each segmentation step with texture analysis used as a final arbitrator between masses and normal structures. The segmentation was evaluated on 168 digitized mammograms and it achieved a performance of 4.4 FPs per image at a 90% true positive (TP) detection fraction and 2.3 FPs per image at an 80% TP detection fraction.[10]

Our approach to mass detection has been to first identify all significant structures within the breast region using a global segmentation technique and then refine the initial object borders using local processing. Finally, we differentiate between true masses and normal structures using morphological and texture information. Our method is therefore different from other detection algorithms that utilize the object shape information for initial detection. The disadvantage of

 0094-2405/99/26(8)/1642/13/$15.00 © 1999 Am. Assoc. Phys. Med.

our combined global and local detection approach is that a large number of normal structures are identified in the initial stage. This can lead to additional FPs if the classification is suboptimal. However, the advantage of this approach is that it can identify difficult masses since the initial detection is not based on shape information. The shape information is still used in the classification stage to reduce FPs.

In this paper, we present an improved version of our two-stage DWCE segmentation approach. This new scheme was designed to both increase specificity and reduce the overall complexity of the segmentation. A primary motivation is to develop a method for eliminating the merging between neighboring structures in the local DWCE processing step and thus improve local segmentation. We introduce an object-based region-growing technique to perform this task. Improved local segmentation serves a number of purposes. First, it improves the morphological and texture information used for FP reduction as well as eliminates the need for the shape-based splitting step. It also enables us to eliminate two morphological FP reduction steps. This significantly reduces the overall complexity of the detection program and should lead to a more practical implementation in a general clinical setting. In this paper, we summarize the intermediate and overall detection performance of the improved mass segmentation algorithm and describe some of its limitations.

II. METHODS

A. Database

The clinical mammograms used in this study were selected from the files of patients who had undergone biopsy at the University of Michigan Hospital. The mammograms were acquired with American College of Radiology (ACR) accredited mammography systems. Kodak MinR/MRE screen/film systems with extended cycle processing were used as the image recorder. The mammography systems have a 0.3-mm focal spot, a molybdenum anode, 0.03-mm thick molybdenum filter, and a 5:1 reciprocating grid. The selection criterion used by the radiologists was simply that a biopsy-proven mass existed on the mammogram. The data set consisted of 253 mammograms from 102 patients, and it included 128 malignant and 125 benign masses. Sixty-three of the malignant and six of the benign masses were judged to be spiculated by a MQSA approved radiologist. The size of the masses ranged from 5 to 29 mm (mean size=12.5 mm), and their visibility ranged from 1 (obvious) to 5 (subtle) (mean=2.1). Figures 1 and 2 show the histograms of mass size and mass visibility for the data set.[13] These distributions characterize the difficulty and diversity of the cases contained in the data set.

The mammograms were digitized with a LUMISYS DIS-1000 laser film scanner with a pixel size of 100 μm and 12 bit gray level resolution. The gray levels were linearly proportional to optical density in the 0.1 to 2.8 optical density unit (O.D.) range. The slope was 0.001 O.D./pixel value. The slope gradually fell off in the 2.8 to 3.5 O.D. range.[10,13] A large pixel value corresponds to a low optical density with this digitizer.

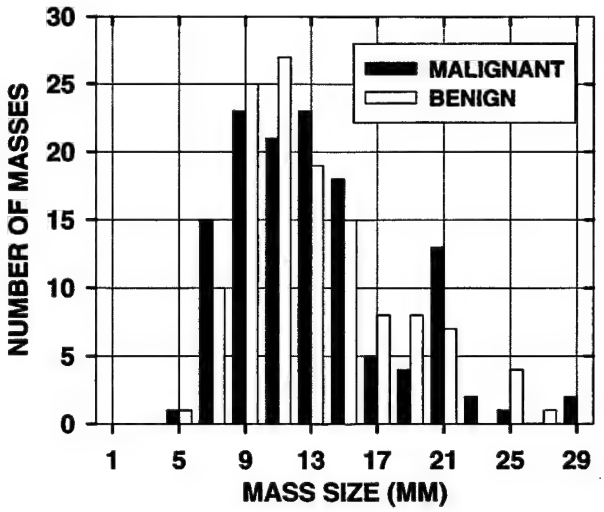


FIG. 1. Histograms of mass size for the 253 masses contained in our data set. Mass sizes were measured as the largest axis of the mass by an experienced breast radiologist.

The location and extent of all the biopsy-proven masses were marked on the original films. The radiologist then identified both the centroid of the lesion and the smallest bounding box containing the entire lesion using an interactive image manipulation tool on a workstation. Both procedures were performed using the original marked film as a guide. The lesion centroid was used to identify TP detections after the morphological FP reduction step. If a segmented object was within 4 mm of the mass centroid, it was considered a TP. All other segmented objects were considered as FPs. The final free-response receiver operating characteristic (FROC) curves following texture-based classification used the more precise mass bounding box for TP identification. A region was considered a TP only when it contained more than 50% of the mass bounding box.

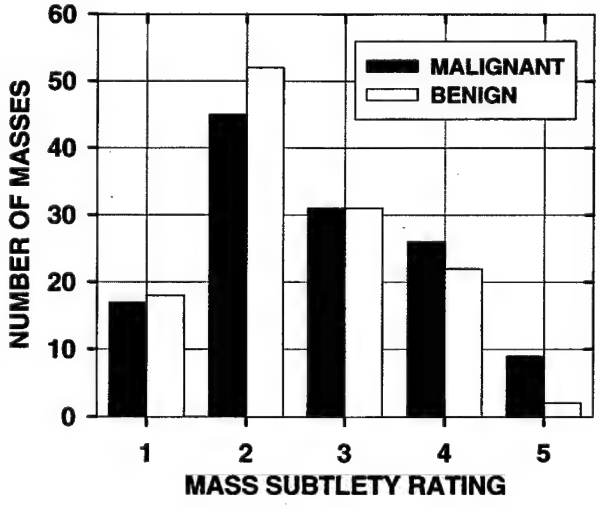


FIG. 2. Histograms of mass subtlety for the 253 masses contained in our data set. Mass subtleties were rated by an experienced breast radiologist from 1 (obvious) to 5 (subtle).

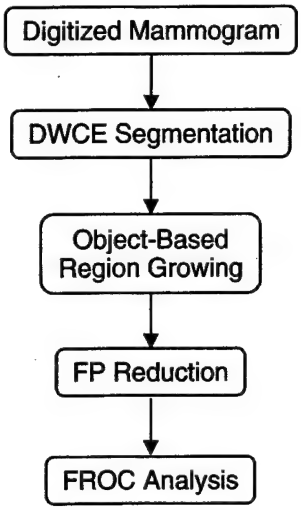


FIG. 3. Block diagram of the breast mass segmentation scheme. A digitized mammogram undergoes DWCE segmentation followed by object-based region growing and then morphological and texture classification. The performance of the segmentation scheme was evaluated by FROC analysis.

B. Density-weighted contrast enhancement segmentation

The block diagram for the proposed detection scheme is shown in Fig. 3. Global DWCE segmentation was used to identify an initial set of breast structures on the digitized mammograms. These objects were then used as seed locations to perform gradient-based region growing. A thorough description of the DWCE technique can be found in the literature.[10,12,14] Briefly, the DWCE technique employs an adaptive filter to enhance the local contrast and thus accentuate mammographic structures in an image. As the term implies, the parameters of the enhancement filter are based on the local density within the image and the filter is applied to the image on a pixel-by-pixel basis. The filter is designed to suppress very low contrast values, to emphasize the low to medium contrast values and to just slightly deemphasize the high contrast values. The effect of suppressing the extremely low contrast values is to reduce bridging between adjacent breast structures. Pixels with low to medium contrast values are enhanced so that more subtle structures can be detected. Finally, the slight deemphasis of the high contrast structures is included to provide a more uniform intensity distribution for detected structures. After contrast enhancement, Laplacian–Gaussian edge detection is applied and all enclosed objects are filled to produce a set of detected structures for the image. The DWCE segmentation is applied to mammograms that have been smoothed and subsampled from their original 100 μm pixel size to an 800 μm pixel resolution.[10] The DWCE stage has been found to be effective in detecting most breast structures including a significant portion of breast masses. However, the DWCE borders usually fall well inside the true borders of an object and a significant number of adjacent structures are merged into single objects. This occurs most frequently when the adjacent breast structures have some tissue overlap.

C. Object-based region-growing segmentation

1. Initial gray-scale region growing

Before gradient-based region growing was applied, an initial set of seed objects was identified. This was accomplished by first identifying all local maxima in the original gray-scale image which occurred within the extent of the DWCE objects. Local maxima were defined using the ultimate erosion technique described by Russ.[15] In simple terms, a pixel was a local maximum if and only if its value was at least as large as all nearest neighbor pixel values. All maxima were identified and grown into larger objects by a simple gray-scale region growing technique as follows. Gaussian smoothing (σ=2.0) was applied to the gray-scale image, and a maximum and a minimum pixel value threshold were specified to select a range of acceptable pixel values. The thresholds were defined as

$$G_i^{\max_1} = 1.01 G_i^{\mathrm{UEP}} \tag{1}$$

and

$$G_i^{\min_1} = 0.99 G_i^{\mathrm{UEP}}, \tag{2}$$

where G_i^{UEP} was the pixel value of the ith maximum and $G_i^{\max_1}$ and $G_i^{\min_1}$ were the maximum and minimum pixel value thresholds, respectively. All pixels within a radius of 20 pixels from a maximum location and with a pixel value inside the defined range were considered to be part of the object. This was repeated for all maxima within an image. Figures 4(a)–4(d) show an original gray-scale image and corresponding images with the DWCE objects, the local maxima, and the gray-scale region-grown objects highlighted. The expanded objects were used as seeds for the gradient-based region growing, described below.

2. Gradient images

A mammogram at 200 μm resolution was used in the gradient-based region-growing stage. The 200 μm resolution image was obtained by averaging 2×2 pixels from the original image. The reduced resolution image had to be smoothed again before gradient filtering because the mammographic tissue produced gradients not only within individual breast structures but also throughout the background portions of the image. Figure 5(b) shows the gradient magnitude image resulting from vertical and horizontal Sobel filtering applied to the 200 μm gray-scale image shown in Fig. 5(a). It clearly demonstrates the large number of gradients throughout the image and the difficulty in applying object-based region growing without additional smoothing. For our application, the smoothing needed to reduce the spurious gradients was accomplished by frequency-weighted Gaussian (FWG) filtering. Frequency-weighted filtering is a technique in which all pixels within the image are split into a base and a residual term. The residual is either positive or negative. This technique produces three subimages from an original image, F, where

$$F = F_F + F_{\mathrm{sub}^+} + F_{\mathrm{sub}^-}. \tag{3}$$

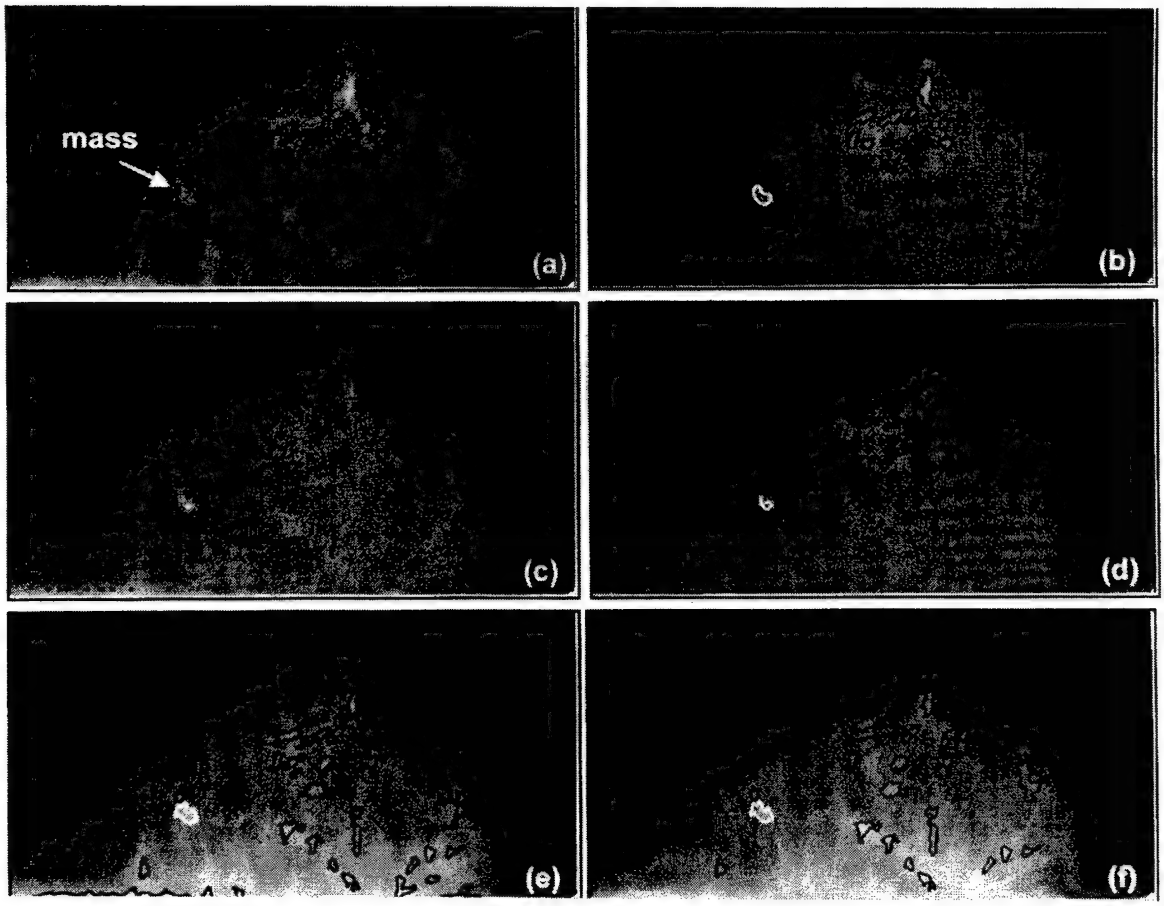


FIG. 4. Objects produced by each segmentation step for a typical mammogram from our data set: (a) the original mammogram with the mass location identified, (b) the DWCE objects, (c) the local maxima, (d) the objects obtained with gray-scale region growing, (e) the objects obtained with gradient-based region growing, and (f) the objects remaining after morphological FP reduction.

The first filter component, F_F, is a filtered version of the original image. In our case, a Gaussian filter, $G(\mu=0,\sigma=10)$, was used. The second and third images are the positive and negative residual images of $F-F_F$, respectively. The F_{sub+} residual is nonzero where the image intensity is larger than the local background and F_{sub-} is nonzero where the image intensity is smaller than the local background. For a particular image pixel, (x,y), the residual images are defined as

$$F_{sub+}(x,y)\equiv\begin{cases}F(x,y)-F_F(x,y), & F(x,y)>F_F(x,y),\\ 0, & \text{otherwise},\end{cases} \tag{4}$$

and

$$F_{sub-}(x,y)\equiv\begin{cases}F(x,y)-F_F(x,y), & F(x,y)<F_F(x,y),\\ 0, & \text{otherwise}.\end{cases} \tag{5}$$

Two FWG filters were designed for sequentially processing the mammograms. The first FWG filtering step reduced the gradients within the breast structures and produced an intermediate image, F_1, which had the form

$$F_1(F)=\tfrac{3}{4}F_F(F)+\tfrac{1}{4}F_{sub+}(F), \tag{6}$$

where the F_F and F_{sub+} images were derived from F, the original 200 μm resolution gray-scale image. A second FWG filtering step was used to eliminate gradients in the breast background. It produced image F_2, which had the form

$$F_2(F_1)=F_{sub+}(F_1), \tag{7}$$

where the F_{sub+} image was derived from image F_1. The result of applying the two FWG filters to the original mammogram in Fig. 5(a) is shown in Fig 5(c). In this image, a significant amount of background has been eliminated and the gradients in the remaining structures have been reduced. Horizontal and vertical Sobel filters[15] were then applied to image F_2 and the magnitude calculated to produce a gradient image as shown in Fig. 5(d). Finally, 5×5 median filtering was used to produce the final gradient image shown in Fig. 5(e). This image was used in the gradient-based region-growing step.

3. Final gradient-based region growing

Each initially grown object (described in Sec. II C 1) was again grown by applying an adaptive technique to the gradient image, F_2, described in Sec. II C 2. The region-growing technique was based on the work of Chang and Li[16] and their adaptive homogeneity test for determining the similarity be-

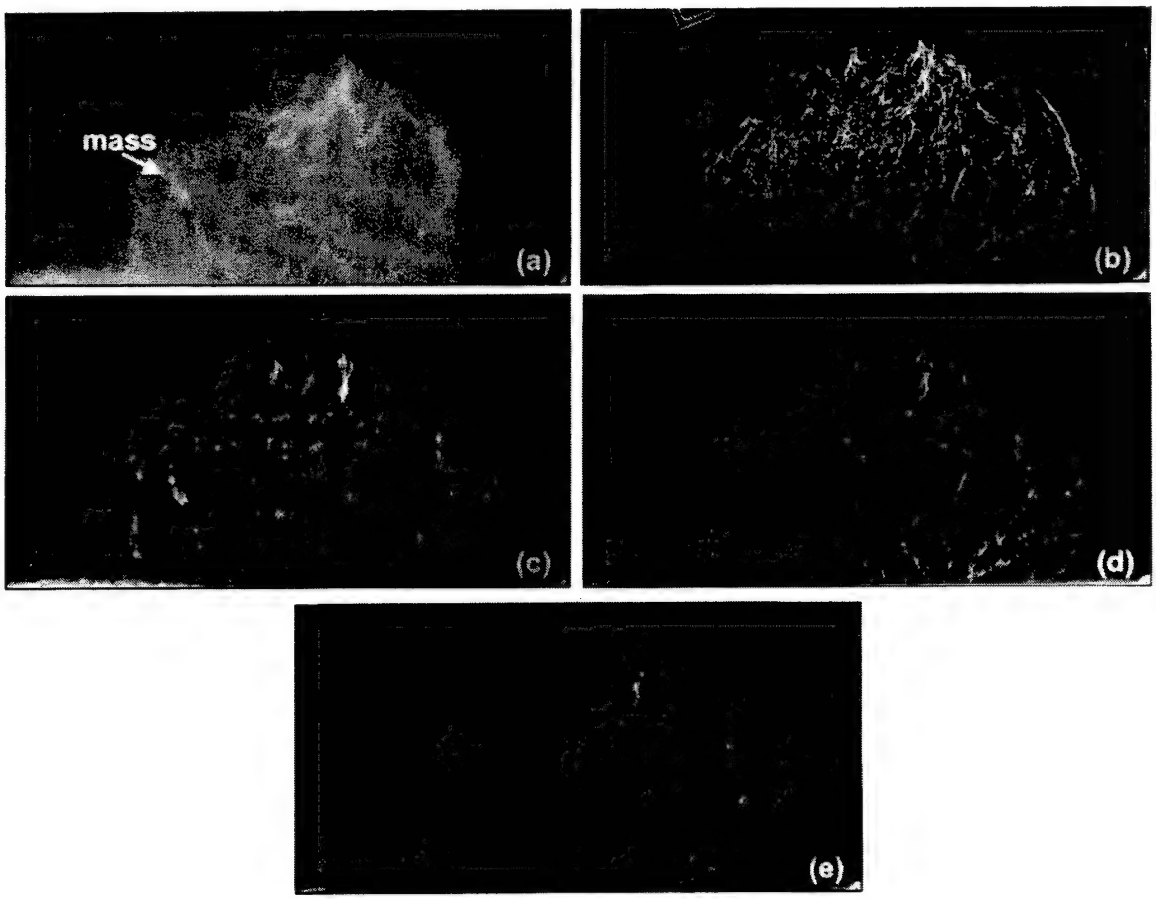


FIG. 5. Processing steps used to define the gradient images: (a) the original mammogram with the mass location identified; (b) the gradient magnitude image obtained from horizontal and vertical Sobel filtering of the original mammogram; (c) the image resulting from FWG filtering of the original mammogram; (d) the gradient magnitude image resulting from horizontal and vertical Sobel filtering of the FWG image; and (e) the image resulting from median filtering of the gradient magnitude image.

tween regions. We have modified this technique to perform object-based region growing. For a mammogram, the corresponding gradient image was smoothed using a Gaussian filter (σ=2.0). A cumulative distribution function (CDF) of pixel values was then calculated from the smoothed gradient image for each object. For each object, the pixel value thresholds were defined as

$$G_{i,0}^{\max_F}=\{g:\mathrm{CDF}_{i,0}(g)=1.0\} \tag{8}$$

and

$$G_{i,0}^{\min_F}=\{g:\mathrm{CDF}_{i,0}(g)=0.0\}, \tag{9}$$

where g was a pixel value and $\mathrm{CDF}_{i,0}(g)$ was the cumulative pixel value distribution within the border of object i and for initial growing iteration 0. The initial growing thresholds simply correspond to the maximum and minimum pixel values within an object. Single-pixel growing was performed on all objects using the thresholds for each individual object to define a range of acceptable pixel values. In this context, single-pixel growing meant growing was limited to only those pixels directly connected to the initial border. Once single-pixel growing was applied to all objects within the image, the thresholds were adjusted and a second iteration of growing was performed. Iterative single-pixel growing was employed to limit the influence of the order that objects were grown within an image. The thresholds used for the ith object during the jth growing iteration were defined as

$$G_{i,j}^{\max_F}=\{g:\mathrm{CDF}_{i,j}(g)=1.0\} \tag{10}$$

and

$$G_{i,j}^{\max_F}=\left\{g:\mathrm{CDF}_{i,j}(g)=\frac{j}{30}\right\}, \tag{11}$$

where $\mathrm{CDF}_{i,j}(g)$ was the cumulative pixel value distribution from the smoothed gradient image within the current borders of object i. Single pixel growing was applied to all objects within the image. This iterative procedure was repeated until no more connected pixels had a value within the appropriately defined range. Note that neighboring objects were not allowed to merge together during this region-growing stage so that growing between adjacent objects stopped with at least a one pixel gap between them. Figures 4(d) and 4(e) show the initial seed objects and the final gradient grown objects for the example shown in Fig. 4(a).

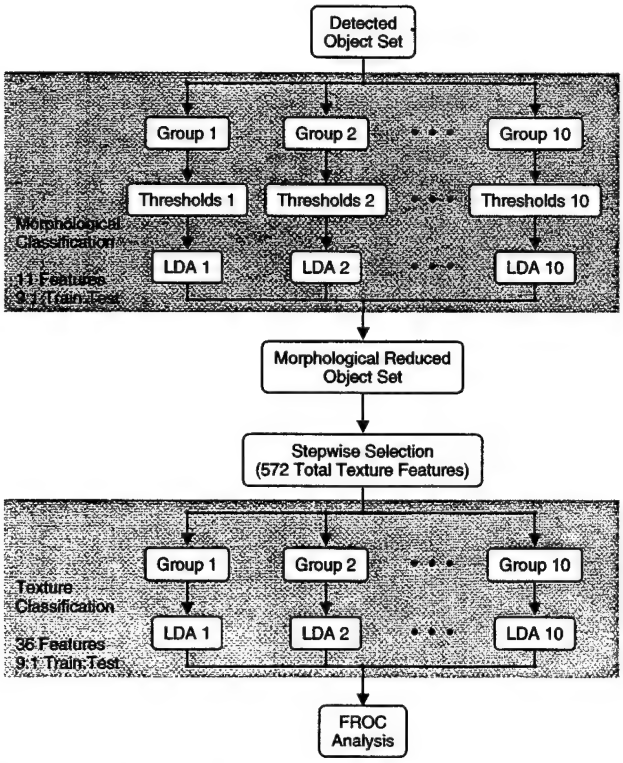


FIG. 6. Flowchart of the FP reduction scheme. The images were separated into ten independent groups. Each group underwent morphological FP reduction with the nine other groups used for classifier training. The reduced objects were recombined and stepwise feature selection was performed. The images were again separated into the ten groups and each group underwent LDA texture classification again using the nine other groups for classifier training. All test scores were then recombined and final FROC analysis was performed.

D. False positive reduction

The DWCE segmentation and region growing do not differentiate masses from normal tissues, therefore, a large number of breast structures were usually detected in each mammogram. Since the shape and texture of mass objects, in general, should be different from those of normal breast structures, a set of features was extracted from each detected object and used to differentiate between the detected structures. The feature set included both morphological and texture features. These features were then used in a sequential classification scheme to reduce the number of FP detections in the mammograms. The sequential application of different classifiers has been found to increase classification accuracy,[17] and it also allows more computationally intensive classifiers to be applied to as few objects as possible. A flow chart depicting the general approach employed for FP reduction is shown in Fig. 6. In this study, morphological classification was initially used to eliminate objects that had shapes significantly different from breast masses. Texture features were then computed for all remaining objects and used with a linear classifier as a final arbiter between masses and normal structures. The following sections describe the major components of the FP reduction scheme.

1. Morphological feature-based FP reduction

The mammograms were partitioned into a number of different groups so that the morphological classifiers could be trained and tested to differentiate masses from normal structures. In this study, the 253 mammograms were randomly partitioned into ten independent groups. Each mammogram was allowed to appear in only one group, and all images from the same patient were grouped together. The goal of the partitioning was to have approximately the same number of images in each group under the given constraints. Classification of the objects within each individual group was performed with a classifier trained using the objects from the nine other image groups. This allowed an approximate 9:1 training-to-test ratio for morphological classification. By rotating the test group through all ten image sets, each mammogram served as a test case once.

Eleven morphological features were used in the initial differentiation of the detected structures. These features included the following object-based measures: number of perimeter pixels, area, perimeter-to-area ratio, circularity, rectangularity, and contrast. In addition, five normalized radial length (NRL) features introduced by Kilday *et al.* were also utilized.[18] They included the NRL mean value, standard deviation, entropy, area ratio, and zero-crossing count. The definition for each morphological feature can be found in the literature.[10] They are also included in Appendix A of this paper.

The morphological features were used as input variables for two different classifiers. A simple threshold classifier was followed by a linear discriminant analysis (LDA) classifier in the morphological FP reduction step. The simple threshold classifier set a maximum and minimum value for each morphological feature based on the maximum and minimum feature values found from the breast masses in the data set. The LDA classification was applied to all objects remaining after threshold classification. The LDA classifier is a linear classifier based on Fisher's discriminant, which is optimal for the two-class, multivariate normal, equal covariance problem.[19,20] The LDA classifier was trained for each training set and applied to the appropriate test set. The LDA classifier produced a single discriminant score for each object in the test set. A threshold was defined as the maximum discriminant score of the masses. This threshold was applied to the test set to further differentiate breast masses for normal structures. The threshold was again based on all masses in the data set to ensure that no mass would be lost during this initial stage. Figure 4(f) shows the results of morphological FP reduction for the example depicted in the figure.

2. Texture feature-based FP reduction

Texture-based classification followed the morphological FP reduction. A large set of multiresolution texture features was extracted for each detected object in the mammogram. Stepwise feature selection was then used to choose the most appropriate set of features for linear classification. The selected features were subsequently used with a LDA classifier to produce a single discriminant score for each detected ob-

TABLE I. The number of detected masses and FPs, the single stage reduction, the mean object area (μ_{Area}), and standard deviation of the object areas (σ_{Area}) for the initial stages in the mass detection scheme. Note texture FP reduction followed the morphological FP reduction stage.

Stage	TPs fraction	FPs/image (initial stages)	Reduction	μ_{Area} (mm^2)	σ_{Area} (mm^2)
DWCE	97%	49.1	...	33.6	66.8
Region growing	97%	45.3	0%	52.4	85.1
Morph. FP reduction	97%	35.5	22%	51.9	52.1

ject. The overall performance of the detection scheme was then evaluated with FROC analysis. The texture-based reduction scheme has been documented in the literature; therefore, this paper will only summarize the important components of the texture analysis and point out any differences from the previously described techniques.[10,21,22]

Regions of interest (ROIs) containing each object remaining after morphological FP reduction were extracted from the 100 μm resolution mammograms. The ROIs had a fixed size of 256×256 pixels and the center of each ROI corresponded to the centroid location of a detected object. The only exception was when the object was located near the border of the breast and a complete 256×256 pixel ROI could not be defined. In this case the ROI was shifted until the appropriate edge coincided with the border of the original mammogram.

Global and local multiresolution texture features, based on the spatial gray level dependence (SGLD) matrix,[23,24] were used in texture analysis.[22] An element of the SGLD matrix, $p_{d,\theta}(i,j)$, is defined as the joint probability that gray levels i and j occur at a given interpixel separation d and direction θ. In this study, 13 texture measures were defined for each SGLD matrix. These measures were correlation, energy, entropy, inertia, inverse difference moment, sum average, sum variance, sum entropy, difference average, difference variance, difference entropy, information measure of correlation 1, and information measure of correlation 2. The definition for all texture measures can be found in the literature[22] and are included in Appendix B of this paper.

The wavelet transform with a four-coefficient Daubechies kernel was used to decompose individual ROIs into different scales. For global texture features, four different wavelet scales, 14 different interpixel distances and 2 different angles were used to produce 28 SGLD matrices. This resulted in 364 global multiresolution texture feature for each ROI. To further describe the information specific to the mass and its surrounding normal tissue, a set of local texture features were calculated for each ROI.[10,22,25] Five rectangular subregions were segmented from each ROI; an object subregion defined by the detected object in the center and four peripheral regions at the corners. Eight SGLD (four interpixel distances and two angles) and a total of 208 local features were calculated from the object subregion and the periphery. They included 104 features in the object region and an additional 104 features defined as the difference between the feature values in the object and the periphery.

In order to improve the generalization of the texture classification, stepwise feature selection was used to select a subset of feature from the pool of 572 global and local features. Feature selection was performed using texture features derived from the ROIs obtained from all 253 images. A total of 40 texture features were selected by stepwise feature selection. Details on the application of stepwise feature selection can be found in our previous publications.[21,26]

At this point in texture classification, the mammograms were again divided into the same ten partitions as described in the morphological FP reduction step. Texture classification was performed on each test group with a trained LDA classifier employing the selected features. The training was based on the texture features derived from the ROIs in the nine other image groups. The test scores within each group were combined with the scores from the other groups to form a complete test set of discriminant scores.

The FROC analysis based on the single set of test scores was used to evaluate the overall performance of the segmentation method.[27,28]

III. RESULTS

The number of TP and FP detections found following the DWCE, region-growing, and morphological FP reduction stages of the segmentation algorithm are summarized in Table I. The DWCE segmentation identified 97% of the breast masses. Table I also includes the reduction percentage, the mean object areas (μ_{Area}) and the standard deviations in the object areas (σ_{Area}) for these initial stages. Table II summarizes the mass type, mass size, mass subtlety, and the

TABLE II. The mass type, mass size, mass subtlety, and mammographic tissue density for the mammograms where the mass was not identified by the initial segmentation. In the table, B dentifies a benign lesion, M identifies a malignant lesion, the subtlety is on a scale of 1 (obvious) to 5 (subtle), and breast density uses the BIRADS density scale of 1 (fatty) to 4 (dense). Both the subtlety and density rankings were performed by an experienced breast radiologist.

Mass no.	Type	Size (mm)	Subtlety	Breast density
1	M	6	4	1
2	B	10	2	1
3	B	14	2	2
4	B	10	2	3
5	B	10	2	3
6	B	14	2	3
7	B	12	4	4
Average		10.9	2.6	2.4

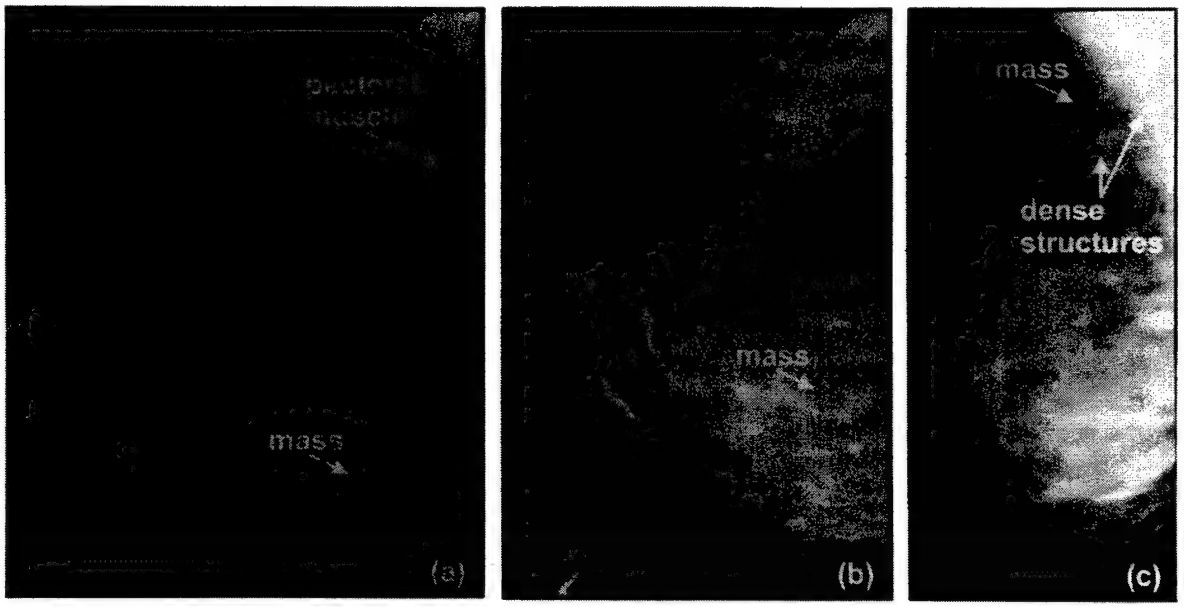


FIG. 7. Examples of masses missed during the initial DWCE segmentation stage: (a) a mammogram with a dense pectoral muscle, fatty breast tissue, and a subtle malignant mass (mass 1 in Table II); (b) a mammogram containing a low contrast benign mass (mass 3 in Table II); and (c) a mammogram with dense structures next to a lower contrast benign mass (mass 4 in Table II).

overall mammographic tissue density for the seven masses missed during the initial DWCE segmentation stage. Figure 7 shows examples of the cases where the mass was missed during the DWCE stage. Figure 8 shows example images with corresponding gradient and object images for cases that had problems during the region-growing stage. This figure contains an example where the mass stopped growing before it reach the correct edge, and an example where the mass was

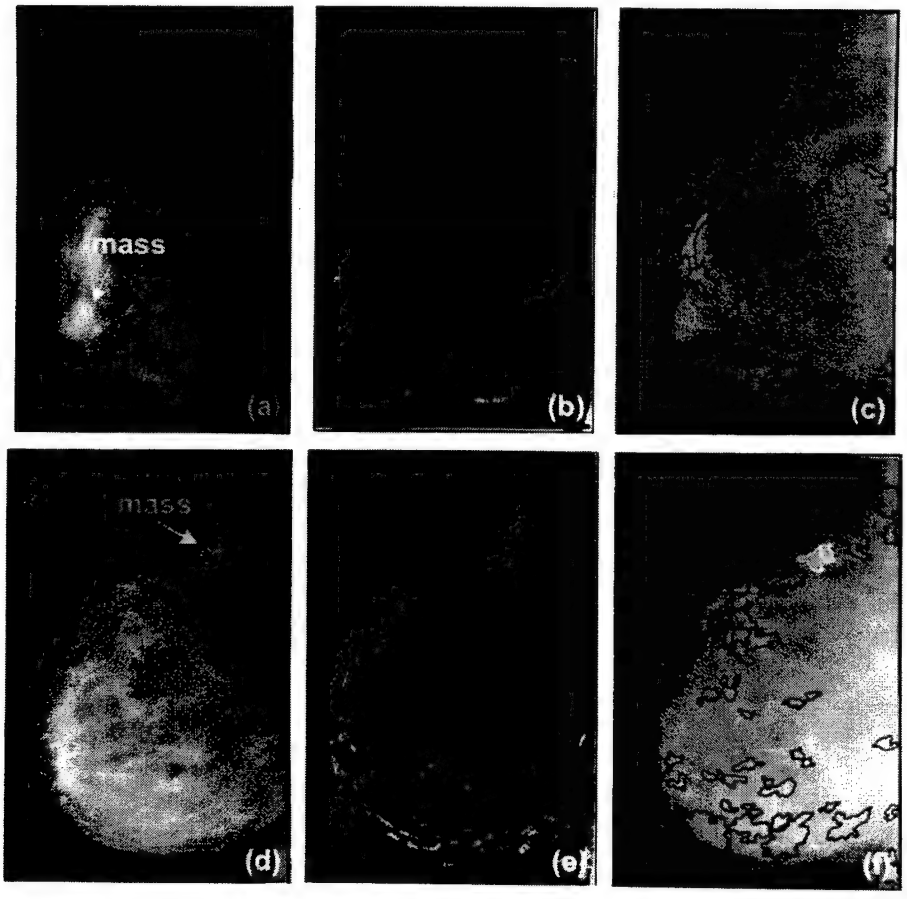


FIG. 8. A mammographic case containing a mass that stopped growing before it reached the correct edge (a)–(c) and a case containing a mass that was split into two pieces during growing (d)–(f). This figure includes (a) and (d) the original mammograms with the mass locations identified, (b) and (e) the corresponding gradient images, and (c) and (f) the final grown objects.

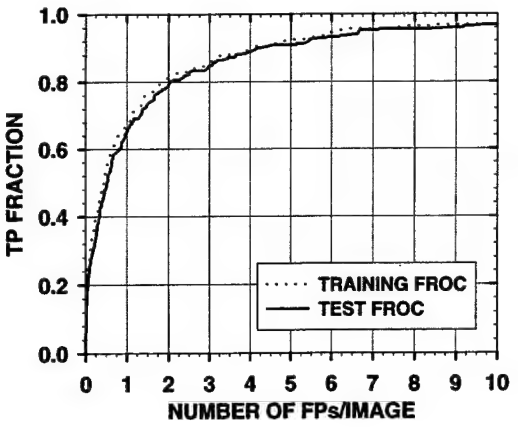


FIG. 9. The training and test FROC curve obtained following LDA classification using 40 selected texture features. The training scores were obtained by averaging the nine training scores from each detected object. The FROC data points were obtained by varying the discriminant decision threshold from the maximum to the minimum value.

split into two pieces during region growing. Finally, Fig. 9 show the FROC training and test performance for the complete segmentation scheme. A summary of the overall performance is given in Tables III and IV for a number of different TP detection fractions. The test performance for the combined DWCE and region-growing segmentation technique at a 90% TP detection level was 4.2 FPs per image and 2.0 FPs per image at an 80% TP level.

IV. DISCUSSION

The purpose of the initial DWCE segmentation stage was to have a method sensitive enough to identify breast masses but which also limited the number of normal structures detected. We have found the DWCE segmentation to be effective in this task. In this study, DWCE segmentation identified 246 of the 253 (97%) masses in the images. Table II summarizes the properties of the masses missed in DWCE segmentation. Masses 1 and 2 were missed because of a dense pectoral muscle visible on the mammogram which overwhelmed all lower-density structures (i.e., both mammograms had BIRADS category 1 breast density). The dense pectoral muscle caused the lower level of the DWCE intensity range to be set so high that lower intensity structures were missed. Figure 7(a) shows the mammogram of the missed malignant mass (mass 1 from Table II). The pectoral muscle is much denser than the mass. This led to the miss. One possible method for eliminating this type of miss may be to identify the pectoral muscle in the mammogram and to apply DWCE segmentation to only the remaining breast region. Mass 3 in Table II was missed because of the small contrast difference between the mass and the background tissue even though the mass was not particularly small or subtle. The mammogram containing this mass is depicted in Fig. 7(b). The remaining masses were missed in mammograms containing denser breast tissue. It was observed that DWCE segmentation had problems detecting masses that were located near much denser normal structures. The dense structures were detected but the masses were missed. Figure 7(c) shows an example of this type of miss. It shows the mammogram containing mass 4 from Table II. Again the dense pectoral muscle may have also hindered detection of the mass in this case. Other than these problems, the DWCE segmentation performed reasonable well as a first stage in mass segmentation. It could identify the majority of the masses while eliminating many of the lower contrast background structures. However, the DWCE segmentation usually underestimated the actual borders of most structures. It also had a tendency to merge the mass with neighboring structures that may have had some tissue overlap with the breast mass. A total of 48 masses had significant merging between the mass and adjacent tissues after DWCE segmentation. This limited the effectiveness of the morphological FP reduction step and limited the localization of the mass during texture-based classification.

The region-growing stage reduced the effects of object merging and significantly increased the size of the initial DWCE objects. This is clearly shown in Table I where the average size of a structure increases from 33.6 mm^2 with DWCE alone to 52.4 mm^2 following region growing. Likewise, a comparison of objects from Figs. 4(b) and 4(e) shows the improvement in border definition following region growing. A combination of gray-scale and gradient-based region growing was used because of the difficulty in stopping gray-scale region growing at the correct edge and the need for large seed objects in gradient-based region growing. The combination approach performed adequately in our detection task and led to an improvement in both morphological and texture-based FP reduction. However, some problems were observed. One problem was that small and low-contrast structures had a tendency to grow into the background and become large regions even though the actual structures were quite small. This did not occur with masses, but it did occur with other breast structures. Another problem was that structures containing internal gradients did not always grow to the correct border, but ended up containing only a section of the true object. This occurred to some mass objects and led to either inaccurate structural information or a mass being split into multiple pieces. Figure 8 shows an example of both incomplete growing and a mass split into pieces during region growing. While these problems reduced the effectiveness of the morphological FP reduction, we have found that the overall benefit of region growing outweights its drawbacks and leads to an improvement in detection accuracy with our segmentation scheme.

The final step in the segmentation was FP reduction. Morphological feature classification was performed first in our reduction scheme. The morphological classification reduced the number of FPs per image from 45.3 to 35.5 as shown in Table I. Following morphological reduction, the average size of the objects was similar to the average size before reduction, but the standard deviation in object size fell from 85.1 mm^2 before reduction to 52.1 mm^2 after reduction. This indicates that morphological reduction eliminated objects that were either much larger or much smaller than the average object size, but had trouble differentiating between TPs and

TABLE III. Summary of the training FROC result depicted in Fig. 9. The table contains the number of FPs per image for different TP fractions along with the percentage of FPs reduced at each TP level relative to the initial value of 19.4 FPs per image. The first entry in the table is the reduction achieved without missing any additional breast masses.

TP fraction	FPs/image	FP reduction
98%	19.4	0%
95%	6.1	69%
90%	4.0	79%
80%	1.9	90%

FPs of similar sizes. Therefore, a classifier that can better differentiate between these similar shaped objects was still necessary. This was achieved, to a large extent, with texture-based feature classification.

A LDA, classifier based on SGLD texture features extracted from ROIs defined by each detected object has proven to be effective in differentiating between similar shaped objects. The training and test FROC performance curves following final texture classification are shown in Fig. 9. In addition, the number of FPs per image for different TP fractions are given in Tables III and IV for the two curves. As discussed in the Methods section, the mammograms were divided into ten independent groups and a 9:1 training-to-test ratio was employed in the classification. Therefore, the test value for an object was its single testing score, and its training value was the average of the scores obtained for the object during training with the nine different training group combinations. The first point to note in Tables III and IV is that the initial TP detection fraction has increased from 97% in Table I to 98% (i.e., 247 total masses were detected). This is due to the change in the definition of a TP with the texture ROIs. The additional mass was detected because in one of the seven mammograms where no object contained the mass centroid, an object ROI overlapped with at least 50% of the mass. The texture classification was able to reduce the number of FPs per image from an initial value of 35.5 to approximately 19 without the loss of any TPs, achieving a 45% reduction. While the number of FPs is still large, it indicates that the more computationally intensive texture classification performs better than morphological reduction. Additional reduction in FPs can be achieved with lower TP detection thresholds. For example, at a 90% TP fraction the FPs decreased to 4.2 per image and at an 80% TP level the FPs decreased to 2.0 per image. Comparing with our previously reported two-stage DWCE edge detection segmentation technique[10] (discussed in Sec. I), we obtained improved performance at all TP levels despite the fact that the data set was increased from 168 to 253 mammograms and two fewer FP reduction stages were used with the new segmentation technique.

TABLE IV. Summary of the test FROC result depicted in Fig. 9. The table contains the number of FPs per image for different TP fractions along with the percentage of FPs reduced at each TP level relative to the initial value of 19.2 FPs per image. The first entry in the table is the reduction achieved without missing any additional breast masses.

TP fraction	FPs/image	FP reduction
98%	19.2	0%
95%	6.7	65%
90%	4.2	78%
80%	2.0	90%

The results presented in this paper do not reflect results from a completely independent test set because the feature selection and the selection of morphological classification thresholds were based on the entire image set. This was necessary to obtain the best possible mass statistics from our limited data set at the intermediate stages of the algorithm. A database is currently being collected so that completely independent testing can be performed using the proposed method.

V. CONCLUSION

We have reported on an improved version of a breast mass detection scheme. The scheme employs DWCE segmentation and object-based region growing. Its overall performance has achieved a 90% TP detection level with 4.2 FPs per image and an 80% TP detection level with 2.0 FPs per image with a diverse database of 253 mammograms. The addition of region growing improved the borders of the detected objects and reduced merging between adjacent or overlapping structures. This improved the morphological information extracted from the detected breast masses and thus the differentiation between masses and normal tissues. The FP reduction was also simplified to a single stage of morphological feature classification and a single stage of SGLD texture feature classification. It is expected that a simplified FP reduction scheme has the potential to generalize better than a more complicated scheme when CAD is implemented in a clinical setting. This breast mass segmentation scheme provided improved FROC performance compared to our previously reported two-stage DWCE technique. Further investigations are under way to improve the region-growing segmentation by analyzing different growing methods that may improve the border definition of the detected structures, as well as to develop new object features that may further differentiate masses from normal structures. Preclinical testing of this algorithm on a large set of independent mammograms will also be conducted.

ACKNOWLEDGMENTS

This work is supported by the Whitaker Foundation (NP), USPHS Grant No. CA 48129, a Career Development Award DAMD 17-96-1-6012 (BS), and research grant DAMD 17-96-1-6254 from the U.S. Army Medical Research and Materiel Command. The content of this publication does not necessarily reflect the position of the government, and no official endorsement of any equipment or product should be inferred.

APPENDIX A: MORPHOLOGICAL FEATURE DEFINITIONS

A set of 11 features is used in morphological FP reduction. Ten of these features are based solely on the binary object defined by the segmentation. The other feature utilizes the original gray scale values inside and surrounding the segmented object. An individual object segmented from image $F(x,y)$ is defined as:

$$F_{\text{obj}_i}(x,y)=\begin{cases}1, & (x,y) \text{ is a pixel in object } i,\\ 0, & \text{otherwise.}\end{cases} \tag{A1}$$

In addition, $F_{BB_i}(x,y)$ defines the pixels contained in the smallest bounding box completely containing object i and $F_{\text{Eqv}_i}(x,y)$ defines the pixels of the circle with the same area as F_{obj_i} and centered at its centroid location. The radius of $F_{\text{Eqv}_i}(x,y)$ is given by

$$r_{\text{Eqv}}=\sqrt{\frac{\text{area }(F_{\text{obj}_i})}{\pi}}. \tag{A2}$$

Five features are based on the normalized radial length (NRL), defined as the Euclidean distance from an object's centroid to each of its edge pixels and normalized relative to the maximum radial length for the object.[18] This results in a NRL vector for each object i given as

$$\mathbf{R}_i=\{r_{i,j}:0\leq j\leq N_e-1\}, \tag{A3}$$

where N_e is the number of edge pixels in the object and $r_{i,j}\leq 1$. The histogram of the normalized radial length is also calculated and is given by

$$\mathbf{P}_i=\{\text{prob}_{i,j}:0\leq j\leq N_h-1\}, \tag{A4}$$

where N_h is the number of bins used in the histogram. Using these basic definitions, the morphological features are defined as follows. Perimeter:

$$\text{Perim}_i=\sum_{\forall x,\forall y} p_i(x,y), \tag{A5}$$

where

$$p_i(x,y)=\begin{cases}1, & F_{\text{obj}_i}(x,y) \text{ is an edge pixel of object } i,\\ 0, & \text{otherwise.}\end{cases}$$

Area:

$$\text{Area}_i=\sum_{\forall x,\forall y} F_{\text{obj}_i}(x,y). \tag{A6}$$

Perimeter-to-area ratio:

$$\text{PAR}_i=\frac{\text{Perim}_i}{\text{Area}_i}. \tag{A7}$$

Circularity:

$$\text{Circ}_i=\frac{\Sigma_{\forall x,\forall y} F_{\text{obj}_i}\cap F_{\text{Eqv}_i}}{\text{Area}_i}. \tag{A8}$$

Rectangularity:

$$\text{Rect}_i=\frac{\text{Area}_i}{\Sigma_{\forall x,\forall y} F_{BB_i}}. \tag{A9}$$

NRL mean:

$$\mu_{\text{NRL}_i}=\frac{1}{N_e}\sum_{j=0}^{N_e-1} r_{i,j}. \tag{A10}$$

NRL standard deviation:

$$\sigma_{\text{NRL}_i}=\sqrt{\frac{1}{N_e}\sum_{j=0}^{N_e-1}(r_{i,j}-\mu_{\text{NRL}_i})^2}. \tag{A11}$$

NRL entropy:

$$E_{\text{NRL}_i}=-\sum_{j=0}^{N_h-1}\text{prob}_{i,j}\cdot\log_2(\text{prob}_{i,j}). \tag{A12}$$

NRL area ratio:

$$\text{AreaR}_i=\left\{\frac{1}{N_e\mu_{\text{NRL}_i}}\sum_{j=0}^{N_e-1}(r_{i,j}-\mu_{\text{NRL}_i}):r_{i,j}>\mu_{\text{NRL}_i}\right\}. \tag{A13}$$

NRL zero-crossing count:

$$\text{ZCC}_i=\sum_{j=0}^{N_e-1} z_{i,j}, \tag{A14}$$

where

$$z_{i,j}=\begin{cases}1, & (r_{i,j-1}>\mu_{\text{NRL}_i})\cap(r_{i,j+1}<\mu_{\text{NRL}_i}),\\ 1, & (r_{i,j-1}<\mu_{\text{NRL}_i})\cap(r_{i,j+1}>\mu_{\text{NRL}_i}),\\ 0, & \text{otherwise.}\end{cases}$$

Contrast:

$$\text{Cont}_i=\frac{g_{\text{in}_i}}{g_{\text{out}_i}}, \tag{A15}$$

where g_{in_i} is the average gray value inside object i and g_{out_i} is the average gray value of the one-pixel wide background surrounding the object.

APPENDIX B: SGLD TEXTURE FEATURE DEFINITIONS

Global and local multiresolution texture features are based on the spatial gray level dependence (SGLD) matrix.[22–24] An element of the SGLD matrix, $p_{d,\theta}(i,j)$, is defined as the joint probability that gray levels i and j occur at a given interpixel separation d and direction θ. In this study, n is defined as the number of gray levels in an image. A total of 13 different texture measures were defined for each SGLD matrix. They were defined as follows.[22]

Energy:

$$E=\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} p_{d,\theta}^2(i,j). \tag{B1}$$

Correlation:

$$R=\frac{\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}(i-\mu_x)(j-\mu_y)p_{d,\theta}(i,j)}{\sigma_x\sigma_y}, \tag{B2}$$

where

$$\mu_x=\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} ip_{d,\theta}(i,j), \tag{B3}$$

$$\mu_y=\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} jp_{d,\theta}(i,j), \tag{B4}$$

$$\sigma_x=\sqrt{\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}(i-\mu_x)^2p_{d,\theta}(i,j)}, \tag{B5}$$

and

$$\sigma_y=\sqrt{\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}(j-\mu_y)^2p_{d,\theta}(i,j)}. \tag{B6}$$

Entropy:

$$H=-\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} p_{d,\theta}(i,j)\log_2(p_{d,\theta}(i,j)). \tag{B7}$$

Inertia:

$$\text{In}=\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}(i-j)^2p_{d,\theta}(i,j). \tag{B8}$$

Inverse difference moment:

$$\text{IDM}=\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\frac{1}{1+(i-j)^2}p_{d,\theta}(i,j). \tag{B9}$$

Sum average:

$$\mu_{x+y}=\sum_{k=0}^{2n-2} kp_{x+y}(k), \tag{B10}$$

where

$$p_{x+y}(k)=\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} p_{d,\theta}(i,j),$$

$$i+j=k \quad \text{and} \quad k=0,\ldots,2n-2. \tag{B11}$$

Sum variance:

$$\sigma^2_{x+y}=\sum_{k=0}^{2n-2}(k-\mu_{x+y})^2p_{x+y}(k). \tag{B12}$$

Sum entropy:

$$H_{x+y}=-\sum_{k=0}^{2n-2} p_{x+y}(k)\log_2(p_{x+y}(k)). \tag{B13}$$

Difference average:

$$\mu_{x-y}=\sum_{l=0}^{n-1} lp_{x-y}(l), \tag{B14}$$

where

$$p_{x-y}(l)=\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} p_{d,\theta}(i,j), \quad |i-j|=l \quad \text{and} \quad l=0,\ldots,n-1. \tag{B15}$$

Difference variance:

$$\sigma^2_{x-y}=\sum_{l=0}^{n-1}(l-\mu_{x-y})^2p_{x-y}(l). \tag{B16}$$

Difference entropy:

$$H_{x-y}=-\sum_{l=0}^{n-1} p_{x-y}(l)\log_2(p_{x-y}(l)). \tag{B17}$$

Information measure of correlation 1:

$$\text{IMC}_1=\frac{H-H_1}{\max\{H_x,H_y\}}. \tag{B18}$$

Information measure of correlation 2:

$$\text{IMC}_2=\sqrt{1-\exp^{-2(H_2-H)}}, \tag{B19}$$

where

$$H_x=-\sum_{i=0}^{n-1} p_x(i)\log_2(p_x(i)), \tag{B20}$$

$$H_y=-\sum_{j=0}^{n-1} p_y(j)\log_2(p_y(j)), \tag{B21}$$

$$H_1=-\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} p_{d,\theta}(i,j)\log_2(p_x(i)p_y(j)) \tag{B22}$$

and

$$H_2=-\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} p_x(i)p_y(j)\log_2(p_x(i)p_y(j)). \tag{B23}$$

[1]L. Tabar *et al.*, "Reduction in mortality from breast cancer after mass screening with mammography," Lancet **1**, 829–832 (1985).
[2]E. L. Thurfjell, K. A. Lernevall, and A. A. S. Taube, "Benefit of independent double reading in a population-based mammography screening program," Radiology **191**, 241–244 (1994).
[3]C. J. Vyborny and M. L. Giger, "Computer vision and artificial intelligence in mammography," AJR, Am. J. Roentgenol. **162**, 699–708 (1994).
[4]N. Karssemeijer and G. te Brake, "Detection of stellate distortions in mammograms," IEEE Trans. Med. Imaging **15**, 611–619 (1996).
[5]H. Kobatake and Y. Yoshinaga, "Detection of spicules on mammogram based on skeleton analysis," IEEE Trans. Med. Imaging **15**, 235–245 (1996).
[6]W. P. Kegelmeyer, J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper, "Computer-aided mammographic screening for spiculated lesions," Radiology **191**, 331–337 (1994).
[7]F. F. Yin, M. L. Giger, K. Doi, C. E. Metz, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: Analysis of bilateral subtraction images." Med. Phys. **18**, 955–963 (1991).
[8]D. Brzakovic, X. M. Luo, and P. Brzakovic, "An approach to automated detection of tumors in mammograms," IEEE Trans. Med. Imaging **9**, 233–241 (1990).
[9]H. D. Li, M. Kallergi, L. P. Clarke, V. K. Jain, and R. A. Clark, "Markov random field for tumor detection in digital ammography," IEEE Trans. Med. Imaging **14**, 565–576 (1995).
[10]N. Petrick, H.-P. Chan, D. Wei, B. Sahiner, M. A. Helvie, and D. D. Adler, "Automated detection of breast masses on mammograms using adaptive contrast enhancement and texture classification," Med. Phys. **23**, 1685–1696 (1996).

[11]H. Kobatake, H. Ron Jin, Y. Yoshinaga, and S. Nawano, "Computer diagnosis of breast cancer by mammogram processing," Radiologia Diagnostica **35**, 29–33 (1994).

[12]N. Petrick, H. P. Chan, B. Sahiner, and D. Wei, "An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection," IEEE Trans. Med. Imaging **15**, 59–67 (1996).

[13]B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," Med. Phys. **25**, 516–526 (1997).

[14]N. Petrick, H. P. Chan, B. Sahiner, M. A. Helvie, M. M. Goodsitt, and D. D. Adler, "Computer-aided breast mass detection: False positive reduction using breast tissue composition," in *Digital Mammography*, edited by K. Doi, M. Giger, R. Nishikawa, and R. Schmidt (Elsevier, New York, 1996).

[15]J. C. Russ, *The Image Processing Handbook* (CRC, Boca Rato, FL, 1992).

[16]Y. L. Chang and X. Li, "Adaptive image region-growing," IEEE Trans. Image Process. **3**, 868–872 (1994).

[17]L. Xu, A. Krzyzak, and C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," IEEE Trans. Syst. Man Cybern. **22**, 418–435 (1992).

[18]J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computer-aided image analysis," IEEE Trans. Med. Imaging **12**, 664–669 (1993).

[19]P. A. Lachenbruch, *Discriminant Analysis* (Hafner, New York, 1975).

[20]R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis* (Wiley, New York, 1973).

[21]D. Wei, H. P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue on digital mammograms: Multiresolution texture analysis," Med. Phys. **22**, 1501–1513 (1995).

[22]D. Wei, H. P. Chan, N. Petrick, B. Sahiner, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "False-positive reduction for detection of masses on digital mammograms: Global and local multiresolution texture analysis," Med. Phys. **24**, 903–914 (1997).

[23]R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," IEEE Trans. Syst. Man Cybern. **SMC-3**, 610–621 (1973).

[24]R. W. Conners, "Towards a set of statistical features which measure visually perceivable qualities of textures," in *Proceedings of the IEEE Conference on Pattern Recognition and Image Processing*, pp. 382–390 (1979).

[25]D. Wei, H. P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, "Multiresolution texture analysis for classification of mass and normal breast tissue on digital mammograms," Proc. SPIE **2434**, 606–611 (1995).

[26]H.-P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space," Phys. Med. Biol. **40**, 857–876 (1995).

[27]D. P. Chakraborty, "Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data," Med. Phys. **16**, 561–568 (1989).

[28]D. P. Chakraborty and L. H. L. Winter, "Free-response methodology, Alternate analysis and a new observer-performance experiment," Radiology **174**, 873–881 (1990).

Stepwise linear discriminant analysis in computer-aided diagnosis: the effect of finite sample size

Berkman Sahiner, Heang-Ping Chan, Nicholas Petrick, Robert F. Wagner*, Lubomir Hadjiiski

Department of Radiology, University of Michigan, Ann Arbor, MI 48109-0904
*Center for Devices and Radiological Health, FDA, Rockville, MD 20857

ABSTRACT

In computer-aided diagnosis (CAD), a frequently-used approach is to first extract several potentially useful features from a data set. Effective features are then selected from this feature space, and a classifier is designed using the selected features. In this study, we investigated the effect of finite sample size on classifier accuracy when classifier design involves feature selection. The feature selection and classifier coefficient estimation stages of classifier design were implemented using stepwise feature selection and Fisher's linear discriminant analysis, respectively. The two classes used in our simulation study were assumed to have multidimensional Gaussian distributions, with a large number of features available for feature selection. We investigated the effect of different covariance matrices and means for the two classes on feature selection performance, and compared two strategies for sample space partitioning for classifier design and testing. Our results indicated that the resubstitution estimate was always optimistically biased, except in cases where too few features were selected by the stepwise procedure. When feature selection was performed using only the design samples, the hold-out estimate was always pessimistically biased. When feature selection was performed using the entire finite sample space, and the data was subsequently partitioned into design and test groups, the hold-out estimates could be pessimistically or optimistically biased, depending on the number of features available for selection, number of available samples, and their statistical distribution. All hold-out estimates exhibited a pessimistic bias when the parameters of the simulation were obtained from texture features extracted from mammograms in a previous study.

Keywords: feature selection, linear discriminant analysis, effects of finite sample size, computer-aided diagnosis

1. INTRODUCTION

A common problem in computer-aided diagnosis (CAD) is the lack of a large number of image samples to design a classifier and to test its performance. The effect of finite sample size on the classification accuracy is therefore an important research topic. In order to treat its specific components, previous studies have mostly ignored the feature selection component of this problem, and assumed that the features used in the classifier were fixed.[1-4] However, in many CAD algorithms, feature selection is a necessary first step. This paper addresses the effect of finite sample size on classification accuracy when the classifier design involves feature selection.

In classifier design, the resubstitution and hold-out estimates are commonly used to assess the accuracy of the classifier. To obtain the resubstitution estimate, the classifier is designed using a number of training samples, and the same samples are then applied to the classifier to yield the distribution of the output decision variable for the training group. The resubstitution performance of the classifier is then measured (e.g., by computing the area under the receiver operating characteristic curve, or by evaluating the probability of misclassification) using this distribution. To obtain the hold-out estimate, the classifier is designed in a similar way, except that an independent set of test samples are applied to the classifier to yield the distribution of the output decision variable for the test group. As the number of training samples increases, both of these estimates approach the true classification accuracy, which is the accuracy of a classifier designed with the full knowledge of the sample distributions. When the training sample size is finite, it is known that, on average, the resubstitution estimate of classifier accuracy is optimistic. In other words, it has a higher expected value than the performance obtained with an infinite design sample set, which is the true classification accuracy. Similarly, on average, the hold-out estimate is pessimistic. When classifier design is limited by the availability of design samples, it is important to obtain a conservative (or pessimistic) performance estimate, which provides a lower bound on the classification accuracy.

In CAD literature, different methods have been used to estimate the classifier accuracy when the classifier design involves feature selection. In a few studies, only the resubstitution estimate was provided.[5] In some studies, the researchers partitioned the samples into training and test groups at the beginning of the study, performed both feature selection and

Part of the SPIE Conference on Image Processing • San Diego, California • February 1999
SPIE Vol. 3661 • 0277-786X/99/$10.00

classifier parameter estimation using the training set, and provided the hold-out performance estimate.[6] Several other studies used a mixture of the two methods: The entire sample space was used as the training set at the feature selection step of classifier design, but once the features were chosen, the hold-out or leave-one-out methods were used to measure the accuracy of the classifier.[7-12] To our knowledge, it has not been reported whether this latter method provides an optimistic or pessimistic estimate of the classifier performance.

This paper describes a simulation study that investigates the effect of finite sample size on classifier accuracy when classifier design involves feature selection. We chose to focus our attention on stepwise feature selection in linear discriminant analysis (stepwise linear discriminant analysis) since this is a simple and common feature selection and classification method. The class distributions were assumed to be multivariate Gaussian. We studied the effect of different covariance matrices and means on feature selection performance. We compared the bias of the classifier when feature selection was performed on the entire sample space, and on the design samples alone. The effects of sample size, number of available features, and parameters of stepwise feature selection on classifier bias were examined.

2. METHODS

To evaluate the effect of sample size on feature selection and classifier bias, we studied the problem of stepwise linear discriminant analysis in two stages. The first stage is stepwise feature selection, and the second stage is the estimation of linear discriminant coefficients for the selected feature subset.

2.1. Stepwise Feature Selection

Stepwise feature selection iteratively enters features into or removes features from the group of selected features based on a feature selection criterion.[13] In our study, we used Wilks' lambda, which is defined as the ratio of within-group sum of squares to the total sum of squares of the discriminant scores, as the feature selection criterion. At the feature entry step of the stepwise algorithm, an F value is computed for each feature based on the ratio of the Wilks' lambda before and after the feature is entered into the pool of already selected features. The feature with the largest F value is entered into the selected feature pool if the F value is larger than a threshold F_{in}. At the feature removal step, the features are tested for removal one at a time from the selected feature pool, the F values are computed, and the feature with the smallest F value is removed from the selected feature pool if the F value is smaller than a threshold F_{out}. The algorithm terminates when no more features can satisfy the criteria for either entry or removal. The number of features selected therefore increases, in general, when F_{in} or F_{out} are reduced.

2.2. Estimation of Linear Discriminant Coefficients

As a by-product of the stepwise feature selection procedure used in our study, the coefficients of a linear classifier that classifies its design samples using the selected features are also computed. However, in this study, the design samples used in the stepwise feature selection step of classifier design may be different from those used in the estimation of classifier coefficients. Therefore, we implemented the stepwise feature selection and the classifier coefficient estimation components of our classification scheme separately.

Let Σ_1 and Σ_2 denote the k-by-k covariance matrices of samples belonging to class 1 and class 2, and let $\mu_1 = (\mu_1(1), \mu_1(2), \ldots, \mu_1(k))$ denote their mean vectors. For an input vector X, the linear discriminant classifier output is defined as

$$h(x) = \frac{1}{2}(\mu_2 - \mu_1)^T \Sigma^{-1} X + \frac{1}{2}(\mu_1^T \Sigma^{-1} \mu_1 - \mu_2^T \Sigma^{-1} \mu_2), \quad (1)$$

where $\Sigma=(\Sigma_1+\Sigma_2)/2$. The linear discriminant classifier is the optimal classifier when the two classes have a multivariate Gaussian distribution with equal covariance matrices.

For the class separation measures considered in this paper (refer to Section 2.3), the constant term $(\mu_1^T \Sigma^{-1} \mu_1 - \mu_2^T \Sigma^{-1} \mu_2)/2$ in Eq. (1) is irrelevant. Therefore, the classifier design can be viewed as the estimation of k parameters of the vector $(\mu_2 - \mu_1)^T \Sigma^{-1}$ using the design samples.

When a finite number of design samples are available, the means and covariances are estimated as the sample means and the sample covariances from the design samples. The substitution of true means and covariances in Eq. (1) by their estimates causes a bias in the accuracy of the classifier. In particular, if the designed classifier is used for the classification of design samples, then the performance is optimistically biased, and if the classifier is used for classifying test samples that are independent from the design samples, then the performance is pessimistically biased.

2.3. Measures of Class Separation

2.3.1. Infinite sample size

When an infinite sample size is available, the class means and covariance matrices can be estimated without bias (i.e., these quantities can be assumed to be known). In this case, we used the Mahalanobis distance $\Delta(\infty)$, or the area $A_z(\infty)$ under the receiver operating characteristic (ROC) curve as measures of classifier accuracy. The infinity sign in parentheses reflects the fact that the distance is computed using the true means and covariance matrices, or, equivalently, using an infinite number of samples.

Assume that the two classes with a multivariate Gaussian distribution with equal covariance matrices have been classified using Eq. (1). Since Eq. (1) is a linear function of the feature vector X, the classifier outputs for class 1 and class 2 will be Gaussian. Let m_1 and m_2 denote means of the classifier output for the normals and the abnormals, respectively, and let s_1^2 and s_2^2 denote the variances for the two classes. With $\Delta(\infty)$ defined as

$$\Delta(\infty)=(\mu_2-\mu_1)^T\Sigma^{-1}(\mu_2-\mu_1), \tag{2}$$

it can easily be shown that

$$m_2-m_1=s_1^2=s_2^2=\Delta(\infty). \tag{3}$$

The quantity $\Delta(\infty)$ is referred to as the Mahalanobis distance between the two classes. It is the Euclidean distance between the two classes, normalized to the common covariance matrix.

In particular, if Σ is an k-by-k diagonal matrix with $\Sigma_{i,i}=\sigma^2(i)$, then

$$\Delta(\infty)=\sum_{i=1}^{k}\delta(i), \tag{4}$$

where

$$\delta(i)=[\mu_2(i)-\mu_1(i)]^2/\sigma^2(i) \tag{5}$$

is the squared signal-to-noise ratio of the difference of the means between the two classes for the i^{th} feature.

Using Eq. (3), and the normality of the classifier outputs, it can be shown that[14]

$$A_z(\infty)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\sqrt{\Delta/2}}e^{-t^2/2}dt \tag{6}$$

2.3.2. Finite sample size

When a finite sample size is available, the means and covariances of the two class distributions were estimated as the sample means and the sample covariances using the training samples, and the classifier outputs for the training and test samples were computed using Eq. (1). The accuracy of the classifier was measured by receiver operating characteristic (ROC) methodology.[15,16] The discriminant scores for samples belonging to class 1 and class 2 were used as decision variables in the LABROC1 program, which provided the ROC curve based on maximum likelihood estimation.

2.4. Simulation conditions

For our simulations, we assumed that the two classes have a multivariate Gaussian distribution with equal covariance matrices, and different means. The number of available features was M=100. We generated a sample size of N_s samples from each class using a random number generator. The sample space was randomly partitioned into N_t training samples and N_s-N_t test samples per class. For a given sample space, we used several different values for N_t in order to study the effect of the design sample size on classification accuracy. In order to reduce the variance of the classification accuracy

estimate, a given sample space was independently partitioned 20 times into N_t training samples and N_s-N_t test samples per class, and the classification accuracy using these 20 partitions was averaged. The procedure described above was referred to as an experiment. For each simulation condition described below, 50 statistically independent experiments were performed, and the results were averaged.

Two methods for feature selection were considered. In the first method, the entire sample space was used for feature selection. In other words, the entire sample space was treated as a training set at the feature selection step of classifier design. Before the coefficient estimation step of classifier design, the sample space was partitioned into training and test groups. The training group was used for classifier coefficient estimation, and the resubstitution and hold-out performances were estimated by applying the training and test groups to the designed classifier, respectively. In the second method, sample set partitioning was performed before feature selection. In other words, both feature selection and coefficient estimation were performed only on the training set.

Case 1: Comparison of correlated and diagonal covariance matrices

Case 1.a

In this simulation condition, the 100X100 covariance matrix Σ was chosen to have a block-diagonal structure

$$\Sigma = \begin{bmatrix} A & 0 & 0 & \cdots & 0 \\ 0 & A & 0 & \cdots & 0 \\ 0 & 0 & A & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & A \end{bmatrix}$$

where the 10X10 matrix A was defined as

$$A = \begin{bmatrix} 1 & 0.8 & 0.8 & \cdots & 0.8 \\ 0.8 & 1 & 0.6 & \ldots & 0.6 \\ 0.8 & 0.6 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0.6 \\ 0.8 & 0.6 & \cdots & 0.6 & 1 \end{bmatrix}$$

and $\Delta\mu(i)$=0.1732 for all i. Using (2), the Mahalanobis distance is computed as $\Delta(\infty)$=3.0, and $A_z(\infty)$=0.89.

Case 1.b

The features in Case 1.a can be transformed into a set of uncorrelated features using a linear transformation, which is called the orthogonalization transformation. The linear orthogonalization transformation is defined by the eigenvector matrix of Σ, so that the covariance matrix after orthogonalization is diagonal. After the transformation, the new covariance matrix turns out to be the identity matrix, and the new mean vector is

$$\Delta\mu(i) = \begin{cases} 0.5477 & \text{if } i \text{ is a multiple of 10} \\ 0 & \text{otherwise} \end{cases}$$

Since a linear transformation will not affect the separability of the two classes, the Mahalanobis distance is the same as in Case 1.a, i.e., $\Delta(\infty)$=3.0.

Case 2: Simulation of a possible condition in CAD

In order to simulate covariance matrices and mean vectors that one may encounter in CAD, we used texture features extracted from patient mammograms in a previous study, which aimed at classifying regions of interest (ROIs) on mammograms as malignant or benign.[7] Ten different spatial gray level dependence (SGLD) texture measures were extracted from each ROI at five different distances and two directions. The number of available features was therefore M=100. The transformations that were applied to the ROI before feature extraction, and the formal definition of SGLD features can be found in the literature.[7,17] The means and covariances for each class were estimated from a database of 249 mammograms.

Case 2.a

In this simulation condition, the two classes were assumed to have a multivariate Gaussian distribution with $\Sigma=(\Sigma_1+\Sigma_2)/2$, where Σ_1 and Σ_2 were estimated from the feature samples for the malignant and benign classes. Since the features have different scales, their variances can vary by as much as a factor of 10^6. Therefore, it is difficult to provide an idea about how the covariance matrix is distributed without listing all the entries of the 100X100 matrix Σ. The correlation matrix, which is normalized so that all diagonal entries are unity, is better suited for this purpose. The absolute value of the correlation matrix is shown as an image in Fig. 1. In this image, small elements of the correlation matrix are displayed as darker pixels, and the diagonal elements, which are unity, are displayed as brighter pixels. From Fig. 2, it is observed that some of the features are highly correlated or anticorrelated. The Mahalanobis distance was computed as $\Delta(\infty)=2.4$, which implied $A_z(\infty)=0.86$.

Case 2.b

To determine the performance of a feature space with equivalent discrimination potential, but independent features, we performed an orthogonalization transformation on the SGLD feature space, as explained previously (Case 1.b).

3. RESULTS

Case 1:

Feature selection from the entire sample space

Figs. 2.a and 2.b plot the area A_z under the ROC curve for the resubstitution and hold-out performance estimates versus the inverse of the number of training samples per class, $1/N_t$, for Case 1.a, and Case 1.b, respectively (number of samples per class N_s=100). The F_{in} value was varied between 0.5 and 1.5, and F_{out} was defined as $F_{out}=max[(F_{in}-1),0]$. Fig. 3 is equivalent to Fig. 2.a, except the number of samples per class was increased from N_s=100 to N_s=500 in this figure.

Case 2:

Feature selection from the entire sample space

The area A_z under the ROC curve for the resubstitution and hold-out performance estimates are plotted versus $1/N_t$ in Figs. 4.a and 4.b for Case 2.a, and Case 2.b, respectively (N_s=100). The F_{in} value was varied between 0.5 and 3.0, and F_{out} was defined as $F_{out}=max[(F_{in}-1),0]$. Fig. 5 is equivalent to Fig. 4.a, except the number of samples per class was increased from N_s=100 to N_s=500 in this figure.

Feature selection from training samples alone

Case 2.a was used as an example. The area A_z under the ROC curves versus $1/N_t$ are plotted for N_s=100 and N_s=500 in Figs. 6 and 7, respectively.

4. DISCUSSION

Fig. 2.b demonstrates the potential disadvantage of performing feature selection using the entire sample space. The best possible test performance with infinite sample size for Case 1 is $A_z(\infty)=0.89$. However, in Fig. 2.b, we observe that some of the "hold-out" estimates were as high as 0.92. These estimates were higher than $A_z(\infty)$ because the hold-out samples were excluded from classifier design only in the parameter estimation stage of the design, and were used as training samples in feature selection. When feature selection is performed using a small sample size, some features that are useless for the general population may appear to be useful for the classification of the small number of samples at hand. This was previously demonstrated in the literature by comparing the probability of misclassification based on either a finite sample set or the entire population subject to the constraint that a given number of features were used for classification.[18] In our study, given a small data set, the variance in Wilks' lambda estimates causes some feature combinations to appear more powerful than they actually are. If the data set is partitioned into training and test groups after feature selection, these feature combinations may provide optimistic hold-out estimates.

The observation made in the previous paragraph about feature selection using the entire sample space is not a general rule, however. Figs. 2.a and 4.a show that one does not always run the risk of obtaining an optimistic bias in the hold-out estimate when the feature selection is performed using the entire sample space. For Case 1, the best possible test performance with an infinite sample size is $A_z(\infty)=0.89$, but the best hold-out estimate in Fig. 2.a is $A_z=0.82$. Similarly, for Case 2, the best possible test performance with infinite sample size is $A_z(\infty)=0.86$, but the best hold-out estimate in Fig. 4.a is $A_z=0.84$. The features in both Case 1.a and Case 2.a were correlated. Case 1.b and Case 2.b were obtained from Case 1.a and Case 2.a by applying a linear orthogonalization transformation to the features so that they become uncorrelated. Figs. 2.b and 4.b show that after this transformation is applied, the hold-out estimates can be optimistically biased for small sample size

(N_s=100). This shows that performing a linear combination of features before stepwise feature selection can have a dramatic influence on its performance. This result is somewhat surprising, because the stepwise procedure is known to select a set of features whose linear combination can effectively separate the classes. However, the orthogonalization transformation in this study is assumed to be known *a priori* (i.e., it is not deduced from the available finite sample size), and is applied to the entire feature space of M features, whereas the stepwise procedure only produces combinations of a subset of these features.

Figs. 6 and 7 demonstrate that when feature selection is performed using the training set alone, the hold-out performance estimate is pessimistically biased. This bias decreases as the number of training samples, N_t, is increased.

When F_{in} and F_{out} values were low, the resubstitution performance estimates were optimistically biased for all the cases studied. Low F_{in} and F_{out} values imply that many features are selected using the stepwise procedure. From previous studies, it is known that a larger number of features in classification leads to larger resubstitution bias.[3] On the other hand, when F_{in} and F_{out} values were very high, the number of selected features could be so low that the resubstitution estimate would be pessimistically biased, as can be observed from Fig. 3 (F_{in}=1.5) and Fig. 4.a (F_{in}=3.0). In all of our simulations, for a given number of training samples N_t, the resubstitution estimate increased monotonically as the number of selected features were increased by decreasing F_{in} and F_{out}.

In contrast to the resubstitution estimate, the hold-out estimate for a given number of training samples did not change monotonically as F_{in} and F_{out} were decreased. This can be observed from Fig. 2.a, where the hold-out estimate for F_{in}=1.5 is larger than all other hold-out estimates with different F_{in} values for N_t=25 ($1/N_t$=0.04). However, for N_t=90 ($1/N_t$=0.011), the hold-out estimate for the same F_{in} value is no longer the largest. In Fig. 2.a, the feature selection was performed using the entire sample space. A similar phenomenon can be observed in Fig. 7, where the feature selection is performed using the training samples alone. This means that for a given number of design samples, there is an optimum value for F_{in} and F_{out} (or the number of selected features) that provides the highest hold-out estimate. This is the well-known peaking phenomenon described in the literature,[19] which can be explained as follows. For a given number of training samples, increasing the number of features in the classification has two opposing effects on the hold-out performance. On the one hand, the new features may provide some new information about the two classes, which tends to increase the hold-out performance. On the other hand, the same features increase the complexity of the classifier, which tends to decrease the hold-out performance. Depending on the balance between how much new information the new features provide and how much the complexity increases, the hold-out performance may increase or decrease when the number of features is increased.

In this study, the number of available features was fixed at M=100. The number of samples per class was N_s=100 in most of the simulations. However, in three of our simulation conditions, we used N_s=500, which meant that the total number of samples was ten times that of available features. The results of these simulations are shown in Fig. 3 for Case 1, and Figs. 5 and 7 for Case 2. Our first observation concerning these figures is that no hold-out estimates in any of these figures are higher than their respective $A_z(\infty)$ values. This suggests that optimistic hold-out estimates may be avoided by increasing the number of available samples, or, possibly, by decreasing the number of features used for feature selection. A second observation is that, compared to other figures in this study, the relationship between the A_z values and $1/N_t$ is closer to a linear relation. This suggests that it may be possible to obtain $A_z(\infty)$ by fitting a line to the A_z vs. $1/N_t$ curves using linear regression, and finding the y-axis intercept. This is similar to the modified Fukunaga and Hayes technique that we discussed previously in the studies of finite sample size effect on classifier bias.

This study examined only the bias of the mean performance estimates, which were obtained by averaging the estimates from fifty experiments as described in Section 2.4. Another important issue in classifier design is the variance of the individual estimates. The variance provides an estimate of the generalizability of the classifier performance to other design and test samples. We previously studied the variance of performance estimates when the classifier design included the estimation of classifier coefficients, but excluded feature selection.[4,20] The extension of our previous studies to include feature selection is an important further research topic.

5. CONCLUSION

In this study, we investigated the finite-sample performance of a linear classifier that included stepwise feature selection as a design step. We compared the resubstitution and hold-out estimates to the true classification accuracy, which is the accuracy of a classifier designed with the full knowledge of the sample distributions. We compared the effect of partitioning the data set into training and test groups before performing feature selection, and after performing feature

selection. When data partitioning was performed before feature selection, the hold-out estimate was always pessimistically biased. When partitioning was performed after feature selection, i.e., the entire sample space was used for feature selection, the hold-out estimates could be pessimistically or optimistically biased, depending on the number of features available for selection, number of available samples, and their statistical distribution. All hold-out estimates exhibited a pessimistic bias when the parameters of the simulation were obtained from correlated texture features extracted from mammograms in our previous study. The understanding of the performance of the classifier designed with different schemes will allow us to utilize a limited sample set efficiently and to avoid an overly optimistic assessment of the classifier

6. ACKNOWLEDGMENTS

This work is supported by a USPHS Grant No. CA 48129 and a USAMRMC grant (DAMD 17-96-1-6254). B. Sahiner and L. Hadjiiski are also supported by Career Development Awards from the USAMRMC (DAMD 17-96-1-6012 and DAMD 17-98-1-8211). N. Petrick is also supported by a grant from the Whitaker Foundation. The authors are grateful to Charles E. Metz, Ph.D., for the LABROC1 programs.

REFERENCES

1. H.-P. Chan, B. Sahiner, R. F. Wagner, N. Petrick, and J. Mossoba, "Effects of sample size on classifier design: Quadratic and neural network classifiers," Proc. SPIE Conf. Medical Imaging 3034, 1102-1113 (1997).

2. R. F. Wagner, H.-P. Chan, J. Mossoba, B. Sahiner, and N. Petrick, "Finite-sample effects and resampling plans: Application to linear classifiers in computer-aided diagnosis," Proc. SPIE Conf. Medical Imaging 3034, 467-477 (1997).

3. H.-P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick, "Effects of sample size on classifier design for computer-aided diagnosis," Proc. SPIE Conf. Medical Imaging 3338, 845-858 (1998).

4. R. F. Wagner, H.-P. Chan, J. Mossoba, B. Sahiner, and N. Petrick, "Components of variance in ROC analysis of CAD_X classifier performance," Proc. SPIE Conf. Medical Imaging 3338, 859-875 (1998).

5. C.-M. Wu, Y.-C. Chen, and K.-S. Hsieh, "Texture feature for classification of ultrasonic liver images," IEEE Transactions on Medical Imaging 11, 141-152 (1992).

6. P. A. Freeborough and N. C. Fox, "MR image texture analysis applied to the diagnosis and tracking of Alzheimer's disease," IEEE Trans. Medical Imaging 17, 475-479 (1998).

7. B. Sahiner, H.-P. Chan, N. Petrick, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," Med. Phys. 25, 516-526 (1998).

8. B. S. Garra, B. H. Krasner, S. C. Horri, S. Ascher, S. K. Mun, and R. K. Zeman, "Improving the distinction between benign and malignant breast lesions: The value of sonographic texture analysis," Ultrasonic Imaging 15, 267-285 (1993).

9. K. G. A. Gilhuijs and M. L. Giger, "Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging," Medical Physics 25, 1647-1654 (1998).

10. M. F. McNitt-Gray, H. K. Huang, and J. W. Sayre, "Feature selection in the pattern classification problem of digital chest radiograph segmentation," IEEE Trans. Medical Imaging 14, 537-547 (1995).

11. Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer," Radiology 187, 81-87 (1993).

12. V. Goldberg, A. Manduca, D. L. Evert, J. J. Gisvold, and J. F. Greenleaf, "Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence," Medical Physics 19, 1475-1481 (1992).

13. N. R. Draper, *Applied regression analysis*, (Wiley, New York, 1998).

14. A. J. Simpson and M. J. Fitter, "What is the best index of detectability," Psychological Bulletin 80, (1973).

15. C. E. Metz, "ROC methodology in radiologic imaging," Invest Radiol 21, 720-733 (1986).

16. C. E. Metz, B. A. Herman, and J.-H. Shen, "Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," Statistics in Medicine 17, 1033-1053 (1998).

17. R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," IEEE Trans. Systems Man Cybernetics SMC-3, 610-621 (1973).

18. S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," IEEE Transactions on Pattern Analysis and Machine Intelligence 13, 252-264 (1991).

19. G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," IEEE Trans. Information Theory 14, 55-63 (1968).

20. R. F. Wagner, H.-P. Chan, B. Sahiner, N. Petrick, and J. T. Mossoba, "Components of variance in ROC analysis of CAD_X classifier performance: Applications of the bootstrap," Proc. SPIE Conf. Medical Imaging 3661, (in print) (1999).

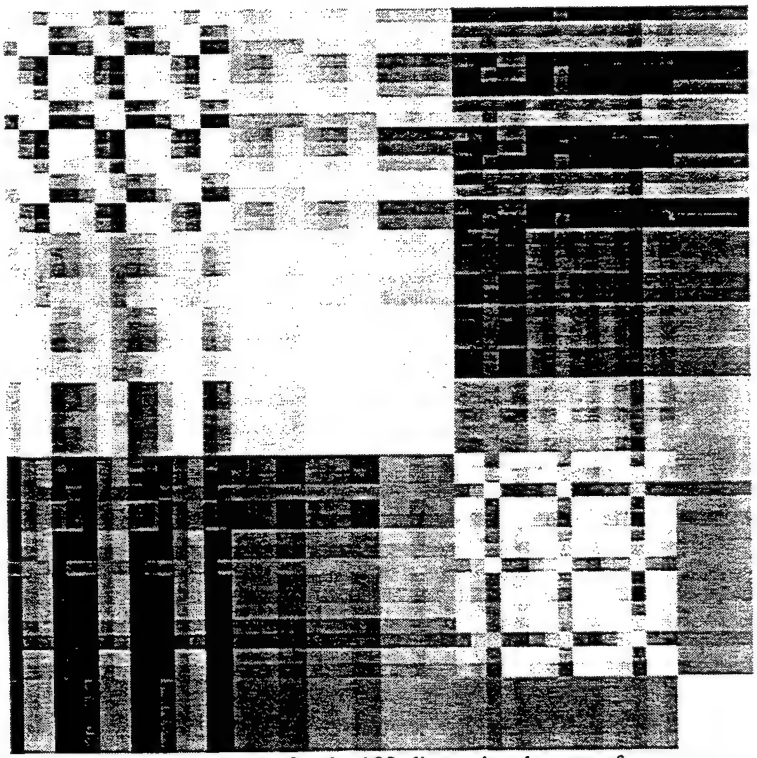


Fig. 1 The absolute value of the correlation matrix for the 100-dimensional texture feature space extracted from 249 mammograms. The covariance matrix corresponding to these features was used in simulation Case 2.a.

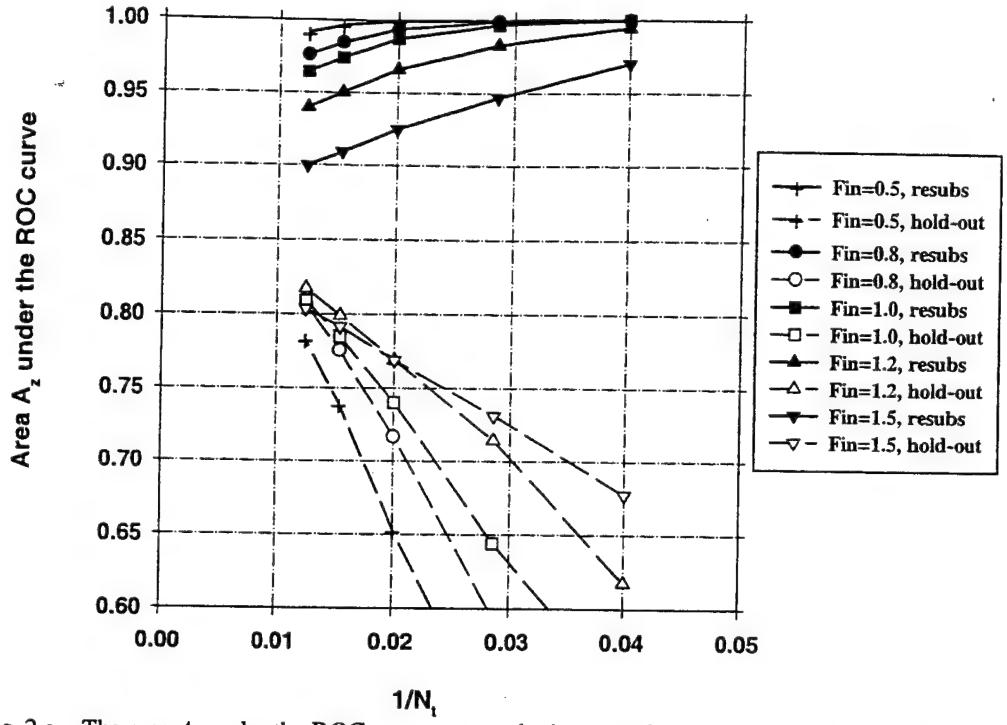


Fig. 2.a The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 1.a, feature selection from the entire sample space of 100 samples/class. Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.89.

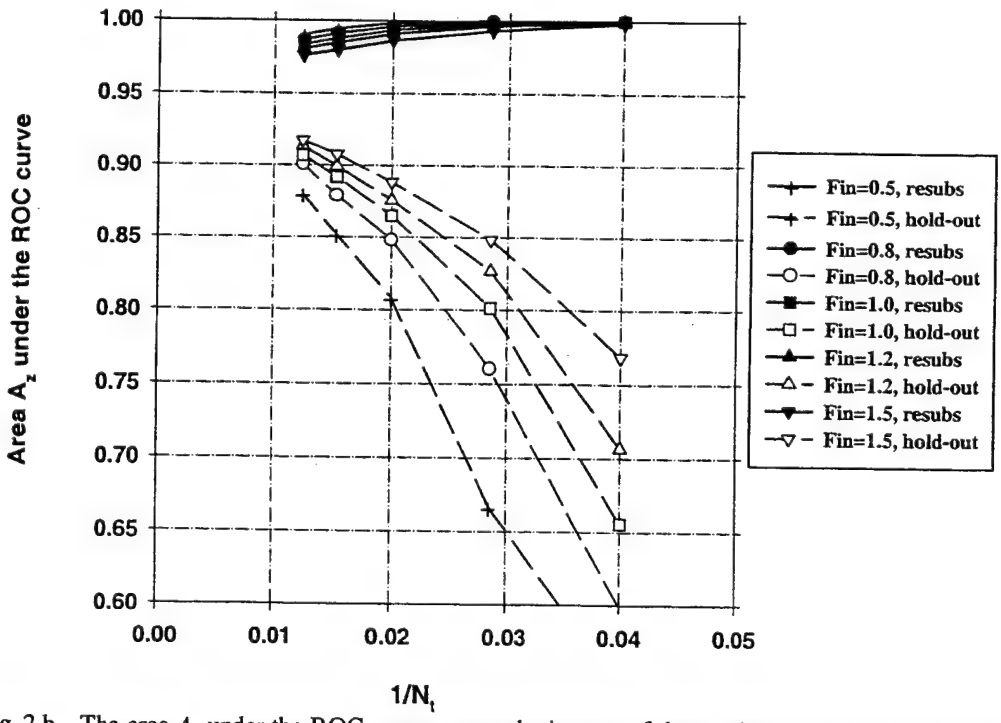


Fig. 2.b The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 1.b, feature selection from the entire sample space of 100 samples/class. Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.89.

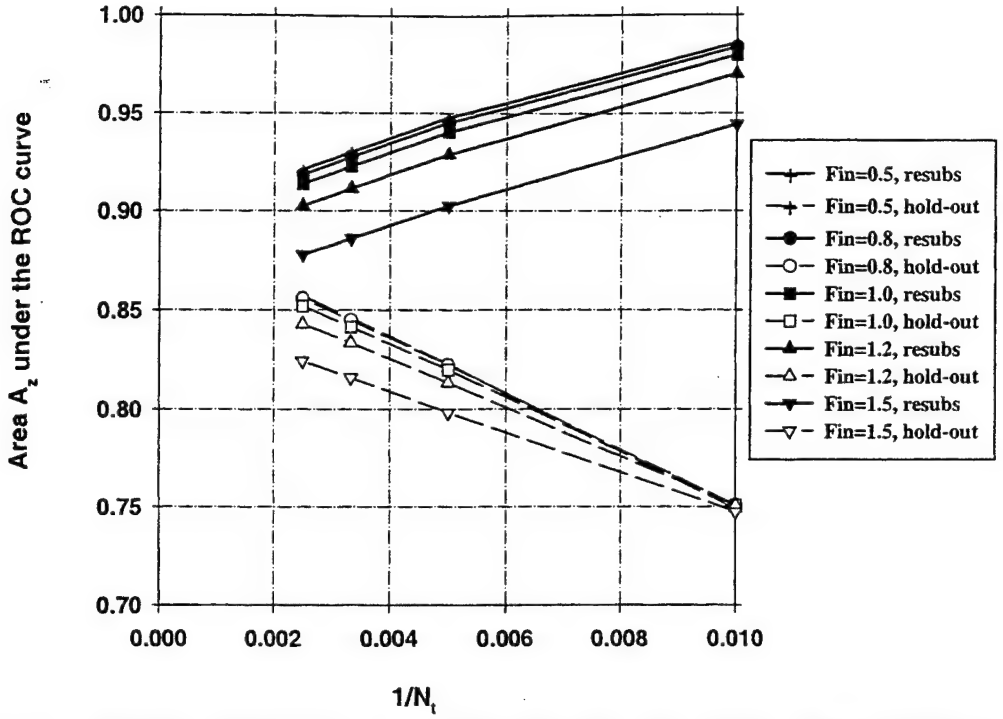


Fig. 3 The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 1.a, feature selection from the entire sample space of 500 samples/class. Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.89.

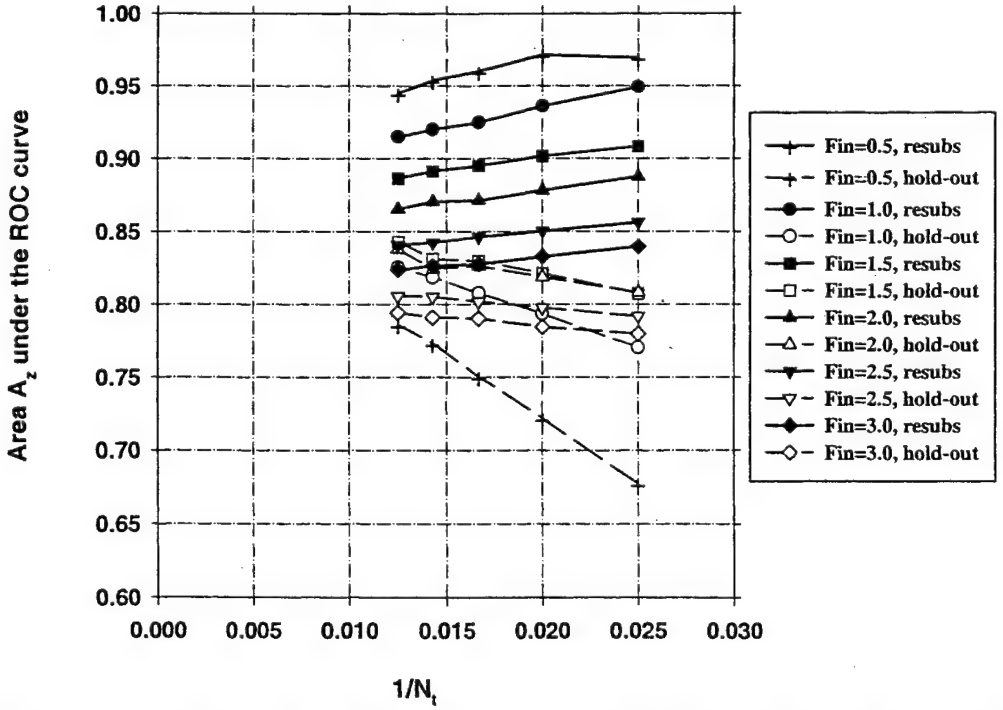


Fig. 4.a The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 2.a, feature selection from the entire sample space of 100 samples/class. Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.86.

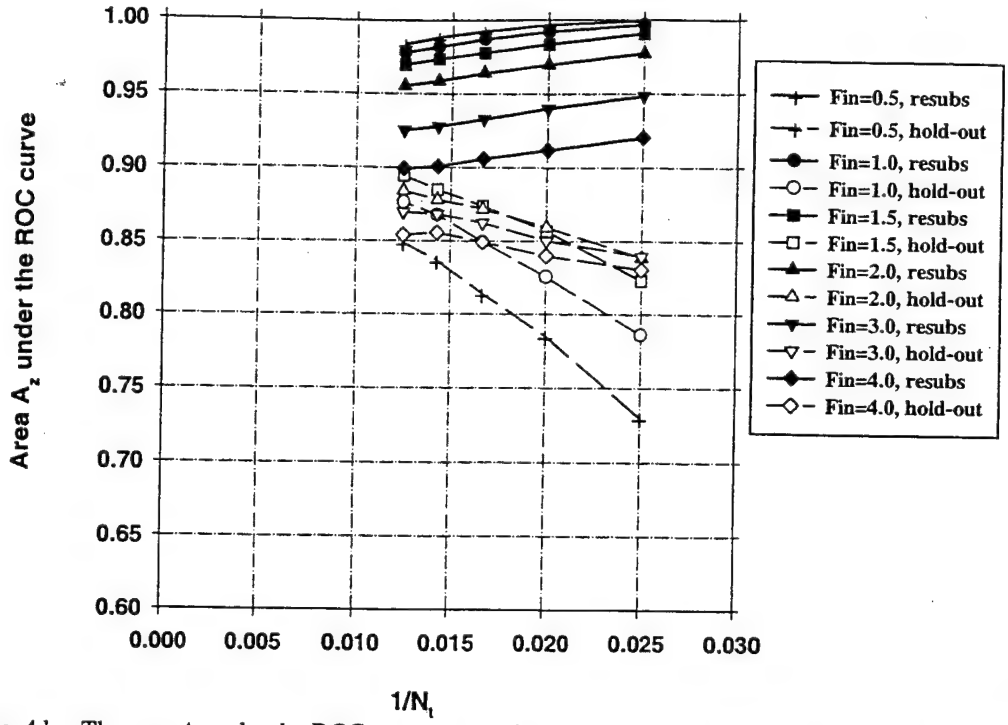


Fig. 4.b The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 2.b, feature selection from the entire sample space of 100 samples/class. Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.86.

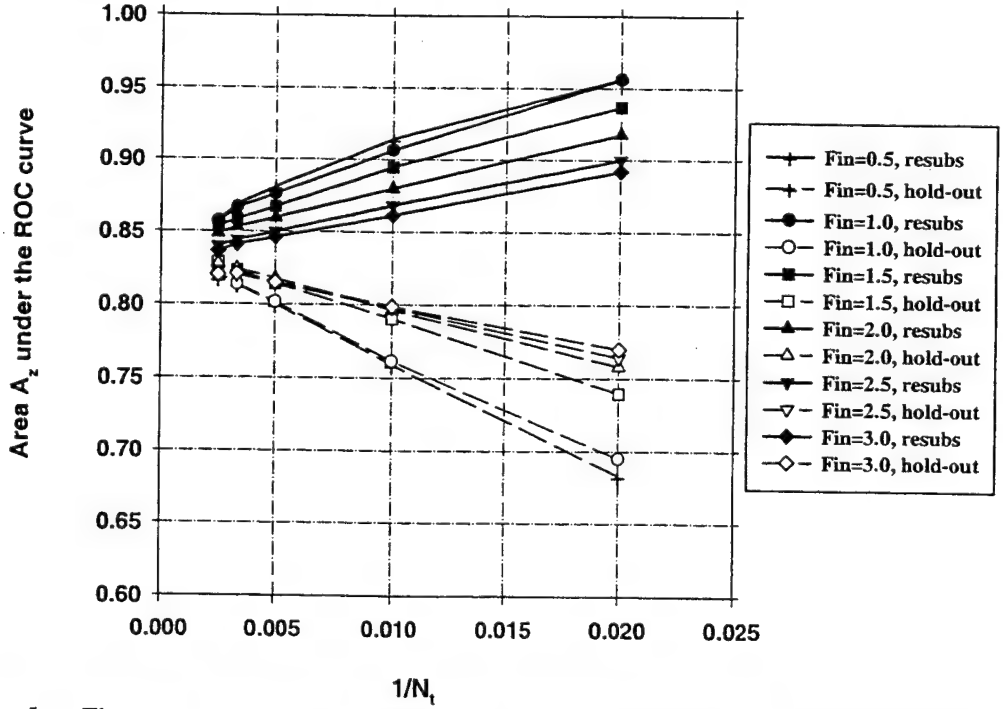


Fig. 5 The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 2.a, feature selection from the entire sample space of 500 samples/class. Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.86.

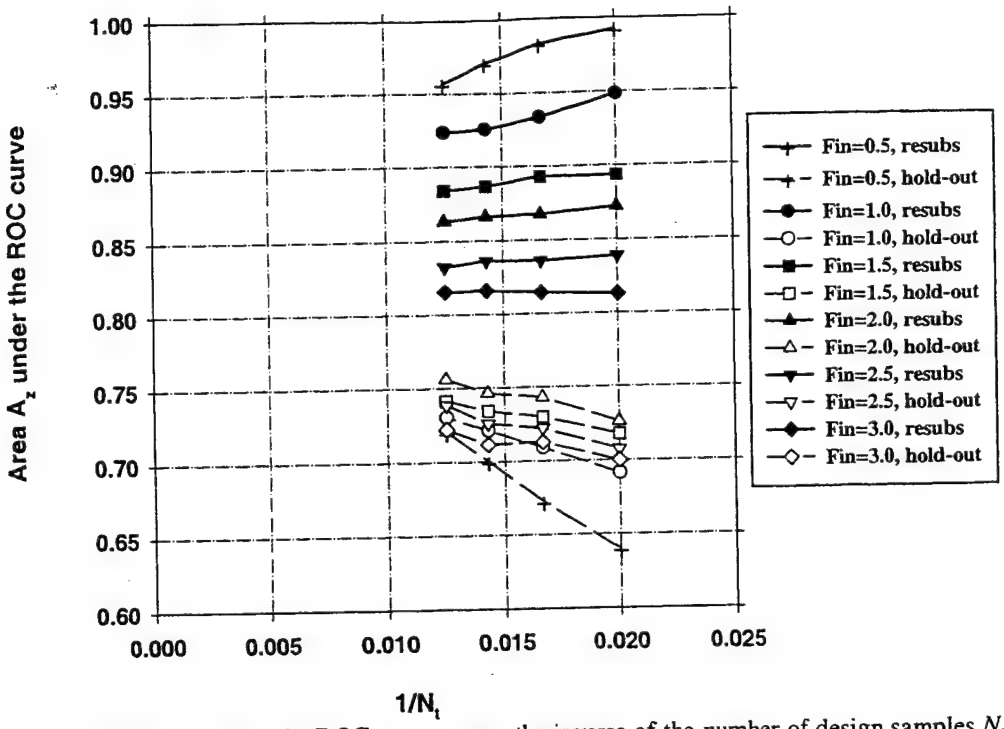


Fig. 6 The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 2.a, feature selection from design samples alone (N_s=100). Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.86.

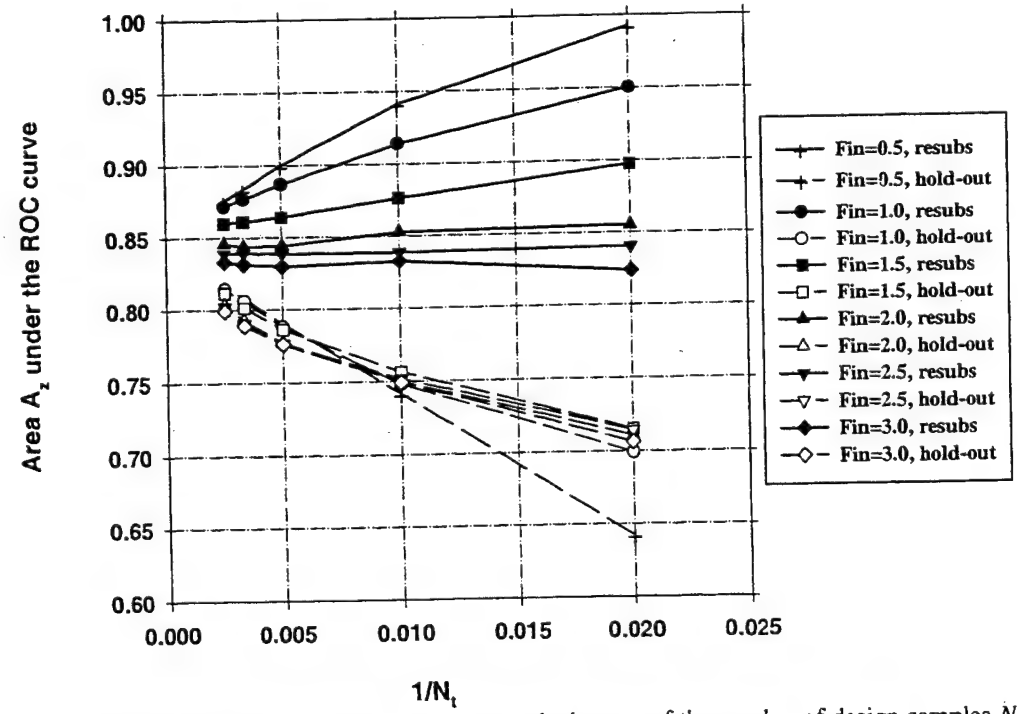


Fig. 7 The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 2.a, feature selection from design samples alone (N_s=500). Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.86.

Hybrid unsupervised-supervised approach for computerized classification of malignant and benign masses on mammograms

Lubomir Hadjiiski, Berkman Sahiner, Heang-Ping Chan, Nicholas Petrick, Mark Helvie

Department of Radiology, The University of Michigan, Ann Arbor, Michigan 48109-0904

ABSTRACT

A hybrid classifier which combines an unsupervised adaptive resonance network (ART2) and a supervised linear discriminant classifier (LDA) was developed for analysis of mammographic masses. Initially the ART2 network separates the masses into different classes based on the similarity of the input feature vectors. The resulting classes are subsequently divided into two groups: (i) classes containing only malignant masses and (ii) classes containing both malignant and benign or only benign masses. All masses belonging to the second group are used to formulate a single LDA model to classify them as malignant and benign. In this approach, the ART2 network identifies the highly suspicious malignant cases and removes them from the training set, thereby facilitating the formulation of the LDA model. In order to examine the utility of this approach, a data set of 348 regions of interest (ROIs) containing biopsy-proven masses (169 benign and 179 malignant) were used. Ten different partitions of training and test groups were randomly generated using 73% of ROIs for training and 27% for testing. Classifier design including feature selection and weight optimization was performed with the training group. The test group was kept independent of the training group. The performance of the hybrid classifier was compared to that of an LDA classifier alone. Receiver Operating Characteristics (ROC) analysis was used to evaluate the accuracy of the classifier. The average area under the ROC curve (A_z) for the hybrid classifier was 0.81 as compared to 0.78 for LDA. The A_z values for the partial areas above a true positive fraction of 0.9 were 0.34 and 0.27 for the hybrid and the LDA classifier, respectively. These results indicate that the hybrid classifier is a promising approach for improving the accuracy of classification in CAD applications.

1. INTRODUCTION

Mammography is the most effective method for detection of early breast cancer[1]. However, the specificity for classification of malignant and benign lesions from mammographic images is relatively low. Clinical studies have shown that the positive predictive value (i.e., ratio of the number of breast cancers found to the total number of biopsies) is only 15% to 30% [2-3]. It is important to increase the positive predictive value without reducing the sensitivity of breast cancer detection. Computer-aided diagnosis (CAD) has the potential to increase the diagnostic accuracy by reducing the false-negative rate while increasing the positive predictive values of mammographic abnormalities.

Classifier design is an important step in the development of a CAD system. A classifier has to be able to merge the available input feature information and make a correct evaluation. Commonly used classifiers for CAD include linear discriminants (LDA)[4] and backpropagation neural networks (BPN)[5] which have been shown to perform well in lesion classification problems[6-9]. These classifiers are generally designed by supervised training. However, these types of classifiers have limitations dealing with the nonlinearities in the data (in case of LDA) and in generalizability when a limited number of training samples are available (especially BPN). Another classification approach is based on unsupervised classifiers, which cluster the data into different classes based on the similarities in the properties of the input feature vectors. Therefore, unsupervised classifiers can be used to analyze the similarities within the data. However, it is difficult to use them as a discriminatory classifier[16,17].

We propose here a hybrid unsupervised/supervised structure to improve classification performance. The design of this structure was inspired by neural information processing principles such as self-organization, decentralization and generalization. It combines the Adaptive Resonance Theory network (ART2)[14,15] and the LDA classifier as a cascade system (ART2LDA). The self-organizing unsupervised ART2 network automatically decomposes the input samples into classes with different properties. The ART2 network performs better compared to conventional clustering techniques in terms of learning speed and discriminatory resolution for the detection of rare events[16,17]. The supervised LDA then classifies the

Part of the SPIE Conference on Image Processing • San Diego, California • February 1999

SPIE Vol. 3661 • 0277-786X/99/$10.00

samples belonging to a subset of classes that have greater similarities. By improving the homogeneity of the samples, the classifier designed for the subset of classes may be more robust.

The ART2LDA design implements both structural and data decomposition. Decomposition is a powerful approach that can reduce the complexity of a problem. Both structural decomposition and data decomposition can improve classification accuracy[10] as well as model accuracy[11]. However, decomposition can also reduce the prediction accuracy due to overfitting the training data. We will demonstrate in this paper that the proposed hybrid structure can deal with the overfitting problem and improve the prediction capabilities of the system.

2. ART2 UNSUPERVISED NEURAL NETWORK

The ART2 is a self organizing system that can simulate human pattern recognition. ART2 was first described by Grossberg[12,13] and a series of further improvements were carried out by Carpenter, Grossberg and co-workers[14,15]. The ART2 network clusters the data into different classes based on the properties of the input feature vectors. The members within a class have similar properties. The process of ART2 network learning is a balance between the plasticity and stability dilemma. Plasticity is the ability of the system to discover and remember important new feature patterns. Stability is the ability of the system to remain unchanged when already known feature patterns with noise are input to the system. The balance between plasticity and stability for the ART2 training algorithm allows fast learning, i.e., rare events can be memorized with a small number of training iterations without forgetting previous events. The more conventional training algorithms such as backpropagation[5] perform slow learning, i.e., they tend to average over occurrences of similar events and require a lot of training iterations.

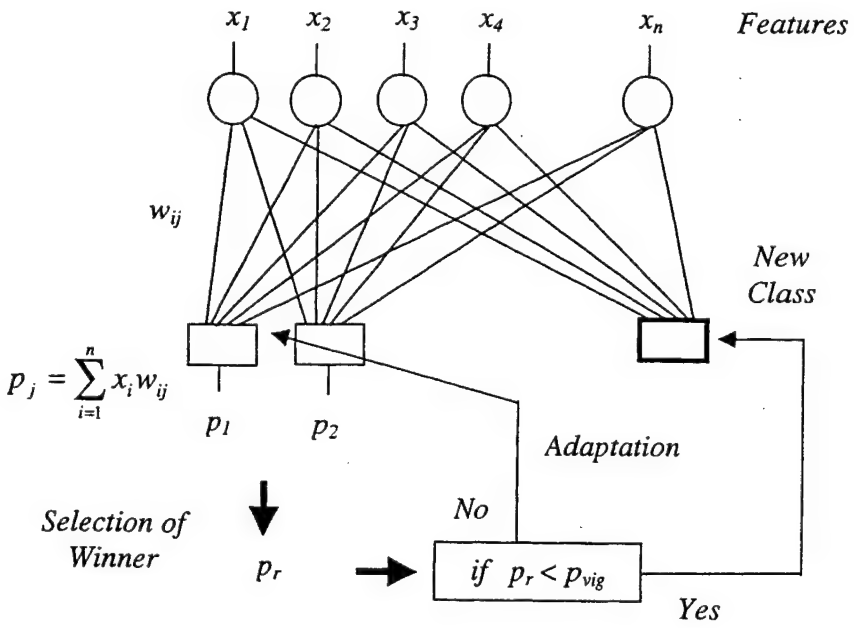


Figure 1. Structure of the ART2 network.

The structure of the ART2 system is shown in Figure 1. It consists of two parts: the ART2 network and the learning stage. Suppose that there are n input features x_i (i=1, ... n) and k classes in the ART2 network. When a new vector is presented to the input of the ART2 network, an activation value p_j for class j is calculated as:

$$p_j = \sum_{i=1}^{n} x_i w_{ij}, \quad j = 1, \ldots, k, \tag{1}$$

where w_{ij} is the connection weight between input i and class j. The activation value is a measure of the membership of the particular input feature vector to class j. The higher the value p_j is, the better the input vector matches class j. The maximum value p_r is selected from all p_j (j = 1, ... ,k) to find the best class match.

Furthermore, in order to balance the contribution to the activation value from all feature components, the input feature values applied to the ART2 system are scaled between zero and one[17]. This normalization will allow detection of similar feature patterns even when the magnitudes of the input feature components are very different.

The learning stage of the ART2 system can influence the weights of the selected class or the complete ART2 network structure by adding a new class. An additional parameter, the vigilance, is used to determine the type of learning[14]. The vigilance parameter p_{vig} is a threshold value that is compared to the maximum activation value p_r. If p_r is larger than p_{vig} then the input vector is considered to belong to class r. The adaptation of the weights connected with class r is performed as follows:

$$w_{ir}^{new} = w_{ir}^{old} + \eta (x_i - w_{ir}^{old}) \quad \text{for } i = 1, \ldots, n, \tag{2}$$

where η is a learning rate. The adaptation of the class r weights (Eq. 2), aims at maximization of the p_r value for the particular input vector. In an iterative manner the weights are adjusted so that the produced activation values for similar input vectors will be maximum only for the class to which they belong and these maximum activation values will be higher than p_{vig}.

If the maximum activation value p_r is smaller than p_{vig}, it is an indication that a novelty has appeared and a new class will be added to the ART2 structure. The new weights connecting the input with the new class ($k+1$) are initialized with the scaled input feature values of this novelty. In this way the activation value p_{k+1} will be maximum ($p_r = p_{k+1}$) and will be higher than p_{vig}, when it is computed for this novelty in further training iterations. The value of the vigilance parameter p_{vig} determines the resolution of ART2. It can be chosen in the range between 0 and 1. If p_{vig} is relatively small, only very different input feature vectors will be distinguished and separated in different classes. If p_{vig} is relatively large the input feature vectors that are more similar will be separated into different classes. The choice of p_{vig} is depends on the particular application.

3. ART2LDA CLASSIFIER

Despite the good performance of ART2 for efficient clustering and detection of novelties, the fast learning approach can cause problems associated with the generalization capability of the system and the correct classification of unknown cases. Supervised classifiers such as linear discriminants or backpropagation neural network classifiers can have better generalization capability than ART2, because they are trained by averaging over similar event occurrences. However, these classifiers do not have the ability to correctly classify rare events.

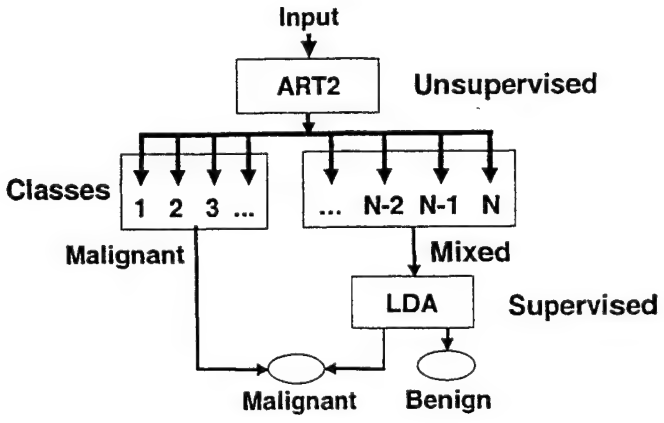


Figure 2. Structure of the ART2LDA classifier.

In order to improve the accuracy and generalization of a classifier, we propose to design a hybrid classifier that combines the unsupervised ART2 network and a supervised LDA classifier. This hybrid classifier (ART2LDA) utilizes the good resolution capability of ART2 and the good generalization capability of LDA. The ART2 network first analyzes the similarity of the sample population and identifies a subpopulation that may be separated from the main population. This will improve the performance of the second-stage LDA if the subpopulation causes the sample population to deviate from a multivariate normal distribution for which LDA is an optimal classifier. Therefore, the ART2 serves as a screening tool to improve the normality of the sample distribution by classifying outlying samples into separate classes.

The structure of the hybrid ART2LDA classifier is shown in Fig. 2. The classes identified by ART2 are labeled to be one of the two types: malignant class or mixed class. A particular class is defined as malignant if it contains only malignant members. It is defined as mixed if it contains both malignant and benign members. The type of a given class is determined based on ART2 classification of the training data set. The ART2 classifies an input sample into either a malignant or a mixed class. Depending on the class type it is determined whether the LDA classifier will be used. If an input sample is classified into a mixed class, the final classification will be obtained based on the LDA classifier, which has been trained by the mixed classes in the training set. However, if an input sample is classified by ART2 into a malignant class then the mass will be considered malignant, without using the LDA classifier. Therefore, in the ART2LDA structure, the ART2 is used both as a classifier and a supervisor.

4. MATERIALS AND METHODS

4.1. Data set

The mammograms used in this study were randomly selected from the files of patients who had undergone biopsy at the University of Michigan. The criterion for inclusion of a mammogram in the data set was that the mammogram contained a biopsy-proven mass. Approximately equal number of malignant and benign masses were included. The data set contained 348 mammograms with a mixture of benign (n=169) and malignant (n=179) masses. The visibility of the masses was rated by a radiologist experienced in breast imaging on a scale of 1 to 10, where the rating of 1 corresponds to the most visible

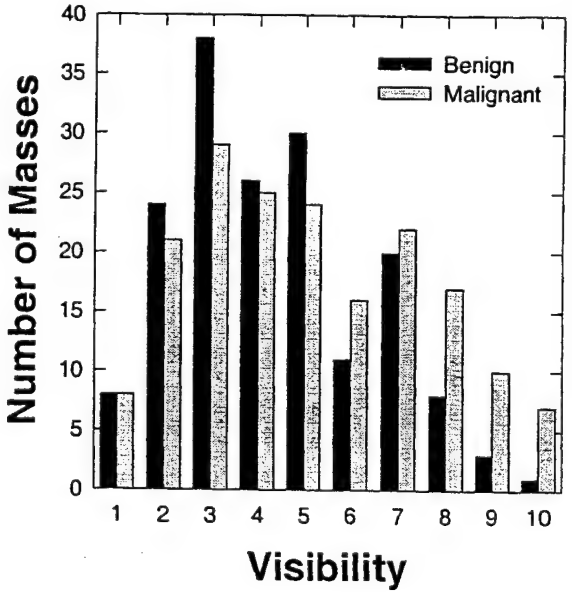


Figure 3. The distribution of the visibility ranking of the masses in the dataset. The ranking was performed by an experienced radiologist. (1: very obvious, 10: very subtle).

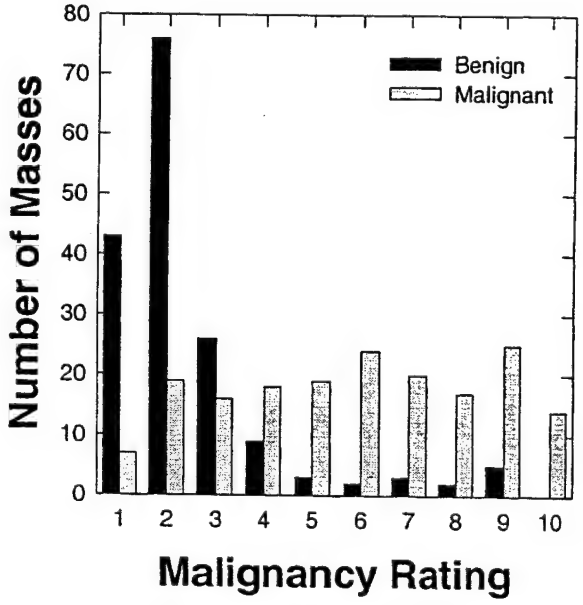


Figure 4. The distribution of the malignancy ranking of the masses in the dataset. The ranking was performed by an experienced radiologist. (1: very likely benign, 10: very likely malignant).

category. The distributions of the visibility rating for both the malignant and benign masses are shown in Fig. 3. The visibility ranged from subtle to obvious for both types of masses. It can be observed that the benign masses tend to be more obvious than the malignant ones. Additionally the likelihood of malignancy for each mass was estimated based on its mammographic appearance. The radiologist rated the likelihood of malignancy on a scale of 1 to 10, where 1 indicated a mass with the most benign appearance. The distribution of the malignancy rating of the masses is shown in Fig. 4.

Three hundred and five of the mammograms were digitized with a LUMISYS DIS-1000 laser scanner at a pixel resolution of 100 μm × 100 μm and 4096 gray levels. The digitizer was calibrated so that gray level values were linearly and inversely proportional to the optical density (OD) within the range of 0.1 to 2.8 OD units, with a slope of –0.001 OD/pixel value. Outside this range, the slope of the calibration curve decreased gradually. The OD range of the digitizer was 0 to 3.5. The remaining 43 mammograms were digitized with a LUMISCAN 85 laser scanner at a pixel resolution of 50 μm × 50 μm and 4096 gray levels. The digitizer was calibrated so that gray level values were linearly and inversely proportional to the OD within the range of 0 to 4 OD units, with a slope of –0.001 OD/pixel value. In order to process the mammograms digitized with these two different digitizers, the images digitized with LUMISCAN 85 digitizer were convolved with a 2×2 box filter and subsampled by a factor of two, resulting in 100 μm images.

In order to validate the prediction abilities of the classifier, the data set was partitioned randomly into training and test subsets. Approximately 73% of the samples have been used for training and 27% for testing. The data set was repartitioned randomly ten times and the training and test results were averaged to reduce their variability.

4.2. Feature extraction

The texture features used in this study were calculated from spatial grey-level dependence (SGLD) matrices[6,7,18] and run-length statistics (RLS) matrices[19]. The SGLD and RLS matrices were computed from the images obtained by the rubber band straightening transform (RBST)[8]. The RBST maps a band of pixels surrounding the mass onto the Cartesian plane (a rectangular region). In the transformed image, the mass border appears approximately as a horizontal edge, and spiculations appear approximately as vertical lines. A complete description of the RBST can be found in the literature[8].

The (i,j)th element of the SGLD matrix is the joint probability that gray levels i and j occur in a direction θ at a distance of d pixels apart in an image. Based on our previous studies[6], a bit depth of eight was used in the SGLD matrix construction, i.e., the four least significant bits of the 12 bit pixel values were discarded. Thirteen texture measures including correlation, energy, difference entropy, inverse difference moment, entropy, sum average, sum entropy, inertia, sum variance, difference average, difference variance and two types of information measure of correlation were used. These measures were extracted from each SGLD matrix at ten different pixel pair distances (d=1, 2, 3, 4, 6, 8, 10, 12, 16 and 20) and in four directions (0°, 45°, 90°, and 135°). Therefore, a total of 520 SGLD features were calculated for each image. The definitions of the texture measures are given in the literature[6-8,18]. These features contain information about image characteristics such as homogeneity, contrast, and the complexity of the image.

RLS texture features were extracted from the vertical and horizontal gradient magnitude images, which were obtained by filtering the RBST image with horizontally or vertically oriented Sobel filters and computing the absolute gradient value of the filtered image. A gray level run is a set of consecutive, collinear pixels in a given direction which have the same gray level value. The run length is the number of pixels in a run[19]. The RLS matrix describes the run length statistics for each gray level in the image. The (i,j)th element of the RLS matrix is the number of times that the gray level i in the image possesses a run length of j in a given direction. In our previous study, it was found experimentally that a bit depth of 5 in the RLS matrix computation could provide good texture characteristics[8].

Five texture measures, namely, short run emphasis, long run emphasis, gray level nonuniformity, run length nonuniformity, and run percentage were extracted from the vertical and horizontal gradient images in two directions, $\theta = 0^\circ$, and $\theta = 90^\circ$. Therefore, a total of 20 RLS features were calculated for each ROI.

A total of 540 features (520 SGLD and 20 RLS) were therefore extracted from each ROI.

4.3. Feature selection

In order to reduce the number of the features and to obtain the best feature set to design a good classifier, feature selection with stepwise linear discriminant analysis[20] was applied. At each step of the stepwise selection procedure one feature is entered or removed from the feature pool by analyzing its effect on the selection criterion. In this study, the Wilks' lambda was used as a selection criterion.

4.4. Performance analysis

To evaluate the classifier performance, the training and test discriminant scores were analyzed using receiver operating characteristic (ROC) methodology. The discriminant scores of the malignant and benign masses were used as decision variables in the LABROC1 program[21], which fit a binormal ROC curve based on maximum likelihood estimation. The classification accuracy was evaluated as the area under the ROC curve, A_z. The discriminant scores of all case samples classified in the two stages of ART2LDA are combined. All masses classified into the malignant group by the ART2 stage were assigned a constant positive discriminant score higher than or equal to the most malignant discriminant score obtained from the LDA classifier.

The performance of ART2LDA was also assessed by estimation of the partial area under the ROC curve ($A_z^{(0.9)}$) at a true positive fraction (TPF) higher than 0.9. The partial $A_z^{(0.9)}$ indicates the performance of the classifier in the high sensitivity (low false negative) region which is most important for cancer detection in clinical practice.

5. RESULTS

In this study, the test subset was kept truly independent from the training subset; only the training subset was used for feature selection and classifier training, and only the test subset was used for classifier validation. In order to validate the prediction abilities of the classifier, ten different partitions of the training and test sets were used and the average classification results were estimated.

Table 1. Number of selected features for the 10 data groups.

Data Group No.	1	2	3	4	5	6	7	8	9	10	Mean
Number of selected features	12	15	13	18	14	14	13	18	14	14	14

For a given partition of training and test sets, feature selection was performed based on the training set. The feature selection results for the ten different training groups are shown in Table 1. The average number of selected features was 14. The selected feature sets contained an average of two RLS features and twelve SGLD features. A different ART2LDA classifier was trained using each training set and the corresponding set of selected features.

5.1. ART2LDA classification results

For the ART2LDA classifier, the number of selected features determines the dimensionality of the input vector of the ART2 classifier and the dimensionality of the LDA classifier. By using different values for the vigilance parameter, ART2 classifiers with different number of classes were obtained. In this study, the vigilance parameter p_{vig} was varied from 0.9 to 0.99, resulting in a range of 10 to 240 classes. The overall performance of the ART2LDA classifier was evaluated for different numbers of ART2 classes because different subset of the samples were separated and classified by ART2. In Fig. 5, the classification results for the ART2LDA are compared to the results from LDA alone for the training and test set partition no. 3. The classification accuracy, A_z, was plotted as a function of the number of ART2 classes. For this training and test set partition, when the number of classes was between 20 and 60, the ART2LDA classifier improved the classification accuracy for the test set in comparison to LDA. As the number of classes increased to greater than 60, the A_z value increased for the training data set, but decreased for the test data set and was lower than that of the LDA alone.

In Table 2 the A_z values of the test set for the 10 corresponding partitions are shown. The average test A_z value is 0.81 for the ART2LDA and 0.78 for LDA alone. For nine of the ten partitions, the A_z value was improved by the hybrid classifier.

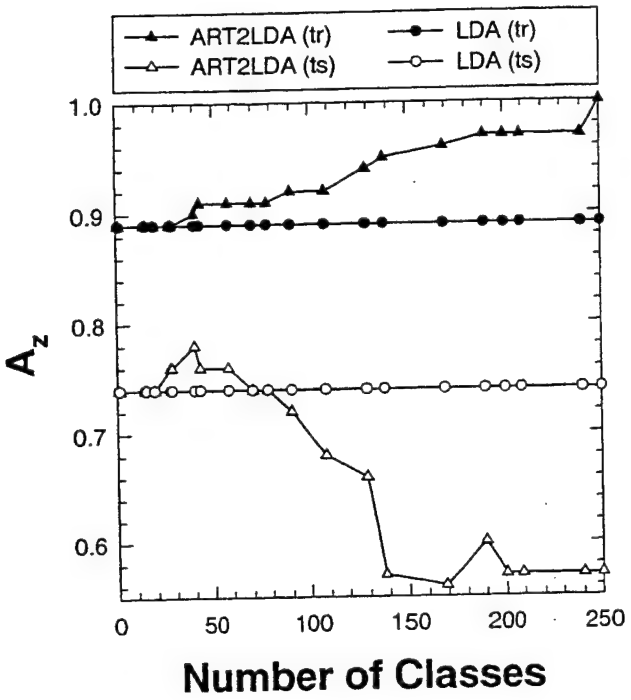


Figure 5. ART2LDA and LDA classification results for training and test sets from data group No.3 as a function of the number of classes generated by ART2.

The performance of ART2LDA was also assessed by estimation of the partial area under the ROC curve $A_z^{(0.9)}$ at a TPF higher than 0.9. In Table 3 the $A_z^{(0.9)}$ values of the test set for the 10 partitions of training and test sets are presented. The average test $A_z^{(0.9)}$ value is 0.34 for the ART2LDA and 0.27 for LDA. For nine of the ten partitions, the $A_z^{(0.9)}$ value was improved at the high sensitivity operating region (TPF>0.9) of the ROC curve.

Table 2. Classifiers performance for the 10 test sets. The A_z values represent the total area under ROC curve.

Data Group No.	LDA	ART2LDA
1	0.77	0.83
2	0.78	0.80
3	0.74	0.78
4	0.77	0.77
5	0.77	0.78
6	0.80	0.83
7	0.80	0.81
8	0.77	0.80
9	0.77	0.80
10	0.86	0.89
Mean	0.78	0.81

Table 3. Classifiers results for the 10 test sets. The A_z values represent the partial area of the ROC curve above the true positive fraction of 0.9 ($A_z^{(0.9)}$).

Data Group No.	LDA	ART2LDA
1	0.14	0.23
2	0.17	0.21
3	0.19	0.32
4	0.19	0.21
5	0.24	0.26
6	0.27	0.38
7	0.32	0.31
8	0.32	0.34
9	0.40	0.49
10	0.44	0.60
Mean	0.27	0.34

6. DISCUSSION

In this paper a new classifier (ART2LDA) is designed and applied to the classification of malignant and benign masses. The results indicate that the ART2LDA classifier has better generalizability than an LDA classifier alone. The ART2 classifier groups the case samples that are different from the main population into separate classes. The minimum number of classes needed to start the clustering of outliers into separate classes depends on how different the outliers are from the rest of the sample population. For the ten different partitions of the training and test sets used in this study, the minimum number varied between 13 and 15 classes. When the number of ART2 classes was less than this minimum number of classes, the ART2 classifier generated only mixed malignant-benign classes and all samples were transferred to the LDA stage. In that case, the ART2LDA was equivalent to the LDA classifier alone. When a higher number of classes was generated, an increased number of cases that may be considered outliers of the general data population was removed (clustered in separate classes). For the ten training sets used in this study, the malignant outliers were gradually removed when the number of classes increased. The training accuracy increased when the number of classes increased and A_z could reach the value of 1.0. However, a large number of ART2 classes led to overfitting the training sample set and poor generalization in the test set. The classification accuracy of ART2 for the test set tended to decrease when the number of classes was greater than about 70. The large number of classes also led to a reduction in the generalizability of the second-stage LDA; the training of LDA with a small number of samples would again result in overfitting the training set, and poor generalizability in the test set. This effect was observed when more than 60 or 70 classes were generated by ART2 (see Fig. 5).

The classification accuracy of ART2LDA increased initially with increased number of classes and then decreased after reaching a maximum. The correct classification of the outliers by the ART2 in combination with an improvement in the classification by the LDA resulted in the increased accuracy. When the number of ART2 classes was further increased, the effects of overfitting by the ART2 and the LDA became dominant and the prediction ability of the ART2LDA decreased. In some cases the second stage LDA prediction was much worse than the ART2. In other cases the ART2 could not generalize well. The generation of a high number of classes is therefore impractical and unnecessary both from computational and methodological point of view.

When the partial area of the ROC curve above the true positive (TP) fraction of 0.9 ($A_z^{(0.9)}$) was considered as a measure of classification accuracy, the advantage of ART2LDA over LDA alone became even more evident. By removing and correctly classifying the outliers the accuracy of the classification is increased at the high sensitivity end of the curve.

We have performed statistical tests with the CLABROC program to estimate the significance in the differences between the A_z values from the ART2LDA and the LDA alone, as well as in the differences in the partial $A_z^{(0.9)}$ from the two classifiers. The statistical tests were performed for each individual data set partition because the correlation among the data sets from the different partitions precludes the use of Student's paired t-test with the ten partitions. We found that the differences in both cases did not reach statistical significance because of the small number of test samples and thus the large standard deviation in the A_z values. However, the consistent improvements in A_z and $A_z^{(0.9)}$ (9 out of 10 data set partitions in both cases) suggest that the improvement was not by chance alone, and that the accuracy of a classification task could be improved by the use of an ART2 network.

An important difference between the classifier designed in this study and many others in the CAD field is the method of feature selection. In several previously published studies[8,22,23] the features were selected from the entire data set first, and then the data set was partitioned into training and test sets. This meant that at the feature selection stage of the classifier design, the entire data set was considered to be a training set. Depending on the distribution of the features and the total number of samples used, the test results in these studies might be optimistically biased[24]. In this study, initially the entire data set was partitioned into training and test sets and then feature selection was performed only on the training set. This method results in a pessimistic estimate of the classifier performance[24] when the training set is small. We therefore expect that the performance will be improved when the classifier designed in this study is trained using a large data set. Since our main purpose in this study was to compare the LDA and ART2LDA classifiers, we did not attempt to quantify how pessimistic our results are in this study.

7. CONCLUSION

A new classifier combining an unsupervised ART2 and a supervised LDA has been designed and applied to the classification of malignant and benign masses. A data set consisting of 348 films (179 malignant and 169 benign) was

randomly partitioned into training and test subsets. Ten different random partitions were generated. For each training set, texture features were extracted and feature selection was performed. An average of fourteen features were selected for each group. Ten hybrid ART2LDA classifiers and ten LDA models alone were trained by using the ten training sets. The average A_z value under the ROC curve for the test sets was better for ART2LDA (A_z=0.81) compared to the LDA alone (A_z=0.78). A greater improvement was obtained when the partial ROC area above a true-positive fraction of 0.9 was considered. The average partial A_z for ART2LDA was 0.34 as compared to 0.27 for LDA. These results indicate that the hybrid classifier is a promising approach for improving the accuracy of classifiers for CAD applications.

ACKNOWLEDGMENTS

This work was supported by a USPHS Grant No. CA 48129, and by U.S. Army Medical Research and Materiel Command (USAMRMC) Grant DAMD 17-96-1-6254. Lubomir Hadjiiski is also supported by a Career Development Award from the USAMRMC (DAMD 17-98-1-8211). Berkman Sahiner is also supported by a Career Development Award from the USAMRMC (DAMD 17-96-1-6012). Nicholas Petrick is also supported by a grant from The Whitaker Foundation. The content of this publication does not necessary reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned in the publication should be inferred. We would like to thank Prof. Stephen Grosberg and Dr. Gail Carpenter for providing us with valuable information as well as for the useful discussions. Additionally we would like to thank Charles E. Metz, Ph.D. for providing the LABROC1 and CLABROC programs.

REFERENCES

1. H. C. Zuckerman, "The role of mammography in the diagnosis of breast canser," in *Breast Canser, Diagnosis and Treatment*, edited by I. M. Ariel and J. B. Cleary (McGraw-Hill, New York, 1987), pp. 152-172.
2. D. B. Kopans, "The positive predictive value of mammography," *Am. J. Roentgenol.* 158, pp. 521-526, 1992.
3. D. D. Adler, and M. A. Helvie, "Mammographic biopsy recommendations," *Curr. Opin. Radiol.* 4, pp. 123-129, 1992.
4. R. O. Duda, and P.E. Hart, *Pattern Classification and Scene Analysis* (Wiley, New York), 1973.
5. D. Rumelhart, G. E. Hinton, and R. J. Williams, in D. E. Rumelhart (ed.), *Parallel and Distributed Processing*, Vol. 1, MIT Press, 1986, pp. 318.
6. H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer- Aided Classification of Mammographic Masses and Normal Tissue: Linear Discriminat Analysis in Texture Feature Space," *Phys. Med. Biol.* 40, pp. 857-876, 1995.
7. D. Wei, H. P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, "Classification of Mass and Normal Breast Tissue on Digital Mammograms: Multiresolution Texture Analysis," *Med. Phys.*, 22, pp. 1501-1513, 1995.
8. B. Sahiner, H. P. Chan, N. Petick, M. A. Helvie, and M. M. Goodsitt, "Computerized Characterization of Masses on Mamograms: The Rubber Band Sraightening Transform and Texture Analysis," *Med. Phys.* 25 (4), pp. 516-526, April 1998.
9. H. P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, K. L. Lam, D. D. Adler and M. M. Goodsitt, "Computerized Classification of Malignant and Benign Microcalsifications on mammograms: Texture analysis using an Artificial Neural Network," *Phys. Med. Biol.* 42, pp. 549-567, 1997.
10. M. Jordan, and R. A. Jacobs, "Hierarchical Mixture of Experts and EM Algorithm," *Neural Computation*, 6, pp. 181-214, 1994.
11. L. Hadjiiski, and P. Hopke, "Design of Large Scale Models Based on Multiple Neural Network Approach," *Intelligent Engineering Systems Through Artificial Neural Networks*, Vol. 7, ASME Press, 1997, pp. 61-66.
12. S. Grossberg, "Adaptive pattern classification and universal recoding. I. Parallel development and coding of neural feature detectors," *Biological Cybernetics*, vol.23, no.3, pp.121-134, 1976.
13. S. Grossberg, "Adaptive pattern classification and universal recoding. II. Feedback, expectation, olfaction, illusions," *Biological Cybernetics*, vol.23, no.4, pp. 187-202, 1976.
14. G. A. Carpenter, and S. Grossberg, "ART 2: self-organization of stable category recognition codes for analog input patterns," *Applied Optics*, vol.26, no.23, 1, pp. 4919-4930, Dec. 1987.
15. G. A. Carpenter, S. Grossberg, and D. B. Rosen, "ART 2-A: an adaptive resonance algorithm for rapid category learning and recognition," *Neural-Networks*, vol.4, no.4, pp. 493-504, 1991.
16. G. A. Carpenter, and N. Markuzon, "ARTMAP-IC and Medical Diagnosis: Instance Counting and Inconsistent Cases," *Neural-Networks*, vol.11, no.2, pp. 323-336, March 1998.

17. Y. Xie, P. K. Hopke, and D. Wienke, "Airborne Particle Classification with a Combination of Chemical Composition and Shape Index Utilizing an Adaptive Resonance Artificial Neural network," *Environmental Science & Technology*, Vol. 28, No. 11, pp. 1921-1928, 1994.
18. R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," *IEEE Trans. Syst. Man Cybern.* 3, pp. 610-621, 1973.
19. M. M. Galloway, "Texture Analysis Using Gray Level Run Length," *Comput. Graph. Image Process.* 4, pp. 172-179, 1975.
20. M. J. Norusis, SPSS Professional Statistics 6.1 (SPSS Inc., Chicago, 1993).
21. C. E. Metz, J. H. Shen, and B. A. Herman, "New Methods for Estimating a Binomial ROC Curve From Continuously Distributed Test Results," *presented at the 1990 Annual Meeting of the American Statistical Association, Anahaim, CA*, 1990.
22. M. F. McNitt-Gray, H. K. Huang, J. W. Sayre, "Feature Selection in the Pattern Classification Problem of Digital Chest Radiograph Segmentation," *IEEE Transaction on Medical Imaging*, Vol. 14, No. 3, pp. 537-547, Sep. 1995.
23. Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated Computerized Classification of Malignant and Benign Masses on Digitized Mammograms," *Acad. Radiol.*, 5, pp. 155-168, 1998.
24. B. Sahiner, H. P. Chan, N. Petrick, R. Wagner, L. Hadjiiski, "The effect of sample size on feature selection in computer-aided diagnosis," *SPIE International Symposium on Medical Imaging*, San Diego, California, February 20-26, 1999., *Proc. SPIE* 3661, (in print).

Digital Mammography: Observer Performance Study of the Effects of Pixel Size on Radiologists' Characterization of Malignant and Benign Microcalcifications

Heang-Ping Chan, Mark A. Helvie, Nicholas Petrick, Berkman Sahiner,
Dorit D. Adler, Caroline E. Blane, Lynn K. Joynt, Chintana Paramagul,
Marilyn A. Roubidoux, Todd E. Wilson, Lubomir M. Hadjiiski, Mitchell M. Goodsitt

Department of Radiology, University of Michigan, Ann Arbor, MI 48109

ABSTRACT

A receiver operating characteristic (ROC) experiment was conducted to evaluate the effects of pixel size on the characterization of mammographic microcalcifications. Digital mammograms were obtained by digitizing screen-film mammograms with a laser film scanner. One hundred twelve two-view mammograms with biopsy-proven microcalcifications were digitized at a pixel size of 35 μm x 35 μm. A region of interest (ROI) containing the microcalcifications was extracted from each image. ROI images with pixel sizes of 70 μm, 105 μm, and 140 μm were derived from the ROI of 35 μm pixel size by averaging 2 x 2, 3 x 3, and 4 x 4 neighboring pixels, respectively. The ROI images were printed on film with a laser imager. Seven MQSA-approved radiologists participated as observers. The likelihood of malignancy of the microcalcifications was rated on a 10-point confidence rating scale and analyzed with ROC methodology. The classification accuracy was quantified by the area, A_z, under the ROC curve. The statistical significance of the differences in the A_z values for different pixel sizes was estimated with the Dorfman-Berbaum-Metz (DBM) method for multi-reader, multi-case ROC data.

It was found that five of the seven radiologists demonstrated a higher classification accuracy with the 70 μm or 105 μm images. The average A_z also showed a higher classification accuracy in the range of 70 to 105 μm pixel size. However, the differences in A_z between different pixel sizes did not achieve statistical significance. The low specificity of image features of microcalcifications and the large interobserver and intraobserver variabilities may have contributed to the relatively weak dependence of classification accuracy on pixel size.

KEY WORDS: Digital mammography, detector, pixel size, microcalcifications, classification, ROC study.

1. INTRODUCTION

X-ray mammography is the most effective diagnostic tool for early detection of breast cancers. However, the image quality of conventional mammography is limited by the contrast sensitivity and dynamic range of screen-film systems. The recent advent of digital detector technology will make digital mammography a clinical reality in the near future. Digital mammography is expected to provide improved image quality that may lead to an improvement in the accuracy of breast cancer diagnosis.

The spatial resolution of current digital detectors is generally lower than that of screen-film systems. Higher resolution digital detectors require smaller pixel sizes. However, development of digital detectors with small pixel sizes is not only technologically demanding, but the requirements for image transmission, archiving, and display also increase rapidly as the matrix size increases. The trade-off between spatial resolution and the cost and efficiency is an important consideration in the development of digital mammography systems.

We have performed an ROC study to evaluate the effects of pixel size on radiologists' characterization of microcalcifications on digitized mammograms. Using seven radiologists experienced in mammography, we compared their classification accuracy of malignant and benign microcalcifications in the pixel size range from 35 μm to 140 μm. The results of the ROC study are discussed in this paper.

2. MATERIALS AND METHODS

2.1 Data Set

Digital mammograms were obtained by digitizing screen-film mammograms with a laser film scanner. One hundred twelve two-view mammograms with biopsy-proven microcalcifications were randomly selected from patient files. The data set included microcalcifications with a range of subtlety. The longest dimension of the cluster ranged from 2 to 18 mm, and a few cases contained diffuse microcalcifications spreading over a large area. All mammograms were digitized at a pixel size of 35 μm x 35 μm. A region of interest (ROI) of 1024 x 1024 pixels containing the microcalcification cluster was extracted from each image. ROI images with pixel sizes of 70 μm, 105 μm, and 140 μm were derived from the ROI of 35 μm pixel size by averaging 2 x 2, 3 x 3, and 4 x 4 neighboring pixels, respectively.

Since viewing images on display monitors can introduce variables that may be difficult to control, we printed the ROI images on film with a laser imager for the observer performance study. To reduce the effects of image size, the ROIs with the three larger pixel sizes (smaller matrix size for the same ROI image) were enlarged to the same printed image size as the 35 μm pixel size image by interpolation. The proper interpolation scheme was chosen by visual comparison of the microcalcification clusters obtained with various interpolation parameters by a radiologist experienced in mammography. The printed ROIs had a size of 84 mm x 84 mm, which corresponded to a pixel pitch of about 82 μm for the laser imager. The printed images were magnified by about 2.3 times, compared with their size on the original screen-film mammograms. However, since radiologists often read microcalcifications with a magnifier, the magnified image should not affect the classification of the microcalcifications. To maintain the same displayed contrast for images of different pixel sizes, the four ROIs of different pixel sizes were printed on the same piece of film. This would minimize the effect of potential fluctuations in the printer calibration and in the development conditions of the laser film on the relative contrast of the printed images.

2.2 Observer Performance Study

Seven MQSA-approved radiologists participated as observers. Each observer read the two-view images in four reading sessions. In each session, one-quarter of the images of each pixel size were read. Each case appeared once and only once in each session. The reading order of the images was randomized for each observer. The likelihood of malignancy of the microcalcifications was rated on a 10-point confidence rating scale and analyzed with ROC methodology[1]. The confidence rating scale was designed by an experienced mammographer and was related to the BI-RADS ratings. A training session was conducted before each reading session to familiarize the observers with the confidence rating scale. The classification accuracy was quantified by the area, A_z, under the ROC curve. The average ROC curve for each reading condition was derived by averaging the slope and intercept parameters obtained from the ROC fitting programs for the individual observers' ROC curves. The statistical significance of the differences in the ROC curves for different pixel sizes was estimated with the Dorfman-Berbaum-Metz (DBM) method for multi-reader, multi-case ROC data[2].

3. RESULTS

The area under the ROC curves for each radiologist observer is shown in Table 1 below. It was found that five of the seven radiologists demonstrated a higher classification accuracy with the 70 μm or 105 μm images than with the 35 μm or 140 μm images. The average A_Z also showed a higher classification accuracy in the pixel size range of 70 to 105 μm. However, the differences in A_Z between different pixel sizes did not achieve statistical significance based on analysis by the DBM method. The intraobserver variability on the likelihood of malignancy ratings was very large. The difference in the ratings for the same film read at different times were as large as 6. The interobserver variablity was also large. The decision threshold for microcalcifications being suspected for malignancy varied over a wide range among the seven radiologists.

Table 1. The area under the ROC curves, A_z, of the individual ROC curves for the seven radiologists (R1, ..., R7) and the area under the average ROC curves obtained from averaging the slope and intercept parameters of the individual ROC curves. The standard deviations of the A_z values are, on average, 0.05.

Pixel Size (μm)	A_z							
	R1	R2	R3	R4	R5	R6	R7	Average
35	0.69	0.62	0.75	0.75	0.65	0.73	0.78	0.71
70	0.73	0.71	0.77	0.80	0.64	0.65	0.77	0.73
105	0.80	0.63	0.74	0.81	0.73	0.60	0.77	0.73
140	0.69	0.64	0.68	0.80	0.68	0.74	0.76	0.71

4. DISCUSSION

The low specificity of image features of microcalcifications and the large interobserver and intraobserver variabilities may have contributed to the relatively weak dependence, if any, of classification accuracy on pixel size. This result is consistent with the finding by Karssemeijer et al.[3] in their ROC study that compared the classification accuracy of microcalcifications on original screen-film mammograms with images digitized at 100 μm pixel size and viewed on a display monitor. The dependence of classification accuracy on pixel size may be further weakened when other patient information is available for making diagnostic decision as in clinical practice. The lower classification accuracy at 35 μm may be caused by the higher image noise level at this small pixel size. Because of the large interobserver and intraobserver variabilities, further studies with a larger data set and a larger number of observers will be needed to determine if the trend observed in this study will achieve statistical significance. In addition, since digitized mammograms and mammograms acquired with digital detectors have different noise, contrast sensitivity, and resolution properties, further investigations are needed to determine if a similar trend holds for digital mammograms.

It may be noted that the current ROC study concentrated on the effect of pixel size on the classification of malignant and benign microcalcifications. Previously we had conducted an ROC study[4] to compare the detectability of subtle microcalcifications on original screen-film mammograms with that on mammograms digitized at 100 μm pixel size using an optical drum scanner. It was found that the detection accuracy of the subtle microcalcifications decreased when radiologists read the 100 μm pixel size digitized images. Our previous study[5] that investigated the detection of microcalcifications by a computer program also indicated a reduction in detectability when the digitization pixel size increased from 35 μm to 140 μm. The results from these experiments indicate that image quality may be more important for the detection task than for the classification task in mammographic imaging.

ACKNOWLEDGMENTS

This work is supported by USPHS Grant CA 48129 and by a grant from the U.S. Army Medical Research and Materiel Command DAMD 17-96-1-6254, Career Development Awards DAMD 17-96-1-6012 (B.S.) and DAMD 17-98-1-8211 (L.H.) from the U.S. Army Medical Research and Materiel Command and a Whitaker Foundation Grant (N. P.). The content of this publication does not necessarily reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned in the publication should be inferred. The authors are grateful to Charles E. Metz, Ph.D., for the LABMRMC program.

REFERENCES

1. C. E. Metz, "Some practical issues of experimental design and data analysis in radiological ROC studies," Invest Radiol 24, 234-245 (1989).

2. D. D. Dorfman, K. S. Berbaum and C. E. Metz, "ROC rating analysis: Generalization to the population of readers and cases with the jackknife method," Invest. Radiol. 27, 723-731 (1992).

3. N. Karssemeijer, J. T. M. Frieling and J. H. C. L. Hendriks, "Spatial resolution in digital mammography," Invest Radiol 28, 413-419 (1993).

4. H. P. Chan, C. J. Vyborny, H. MacMahon, C. E. Metz, K. Doi and E. A. Sickles, "Digital mammography: ROC studies of the effects of pixel size and unsharp-mask filtering on the detection of subtle microcalcifications," Invest Radiol 22, 581-589 (1987).

5. H. P. Chan, L. T. Niklason, D. M. Ikeda, K. L. Lam and D. D. Adler, "Digitization requirements in mammography: Effects on computer-aided detection of microcalcifications," Med Phys 21, 1203-1211 (1994).

Monte Carlo Validation of a Multireader Method for Receiver Operating Characteristic Discrete Rating Data: Split Plot Experimental Design

Donald D. Dorfman,[a,b] Kevin S. Berbaum,[a,b] Russell V. Lenth,[c] Yeh-Fong Chen [c]

[a]Department of Radiology, [b]Department of Psychology, and [c]Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, Iowa.

ABSTRACT

The major purpose of this paper was to evaluate the Dorfman/Berbaum/Metz[1] (DBM) method for analyzing multireader receiver operating characteristic (ROC) discrete rating data on reader split-plot and case split-plot designs. It is not always appropriate or practical for readers to interpret imaging studies of the same patients in all modalities. In split plot designs, either a different sample of readers is assigned to each modality or a different sample of cases is assigned to each modality. For each type of split-plot design, a series of null-case Monte Carlo simulations were conducted. The results suggest that the DBM method provides trustworthy alpha levels with discrete ratings when ROC area is not too large, and case and reader sample sizes are not too small. In other situations, the test tends to be somewhat conservative. Our Monte Carlo simulations show that the DBM multireader method can be validly extended to the reader-split and case-split plot designs.

Keywords: Receiver operating characteristic curve (ROC); Diagnostic radiology; Decision theory, Analysis of variance.

1. INTRODUCTION

The major purpose of this paper was to evaluate the Dorfman/Berbaum/Metz[1] (DBM) method for analyzing multireader receiver operating characteristic (ROC) discrete rating data on reader-split and case-split designs. The method involves analysis of variance of pseudovalues computed by the Quenouille-Tukey jackknife. The basic data for the analysis are pseudovalues of ROC parameters computed by jackknifing cases separately for each observer. The problem of multireader ROC analysis is of considerable importance in the evaluation of diagnostic imaging systems. Recently, Obuchowski and Zepp[2] reviewed the major papers on prospective studies of image interpretation published in the *American Journal of Roentgenology* in the first four months of the years 1990 and 1995. They discovered an important trend: "*In the 1990 literature, we noted eight multiple-reader and 18 single-reader studies; in contrast, in the 1995 literature, we found 29 multiple-reader and eight single-reader studies. This trend reflects an increased awareness of the importance of multiple-reader studies.*"

A principal advantage of the fully crossed factorial design analyzed by Dorfman, Berbaum, Metz[1] is that it provides good precision for comparing modalities because between-reader variability is excluded from the experimental error. In the factorial design, only within-reader variation enters the experimental error, since any two modalities can be compared directly for each reader. It is not always appropriate or practical for readers to interpret imaging studies of the same patients in all modalities. For instance, when it is rare for readers to be expert in both modalities being compared, it is better to assign different readers to each modality. Also, when an interpretation of a case in one modality can affect interpretation of that case in another modality, it is appropriate for each same reader to interpret different cases in each

Part of the SPIE Conference on Image Perception and Performance
San Diego, California • February 1999
SPIE Vol. 3663 • 0277-786X/99/$10.00

modality. Under these circumstances, split-plot designs offer solutions. In split plot designs, either a different sample of readers is assigned to each modality or a different sample of cases is assigned to each modality. In our adaptation of the reader split-plot design, it is assumed that readers, called blocks, are a random sample from some population and that each reader is tested under one of the two modalities. To facilitate comparison between the factorial and reader split-plot designs, we assume that there are *n* different readers in each modality, whereas in the factorial design, the same *n* readers are tested in both modalities. Similarly, in the case split-plot design we assume that there are *c* different cases in each modality, whereas in the factorial design the same *c* cases are tested in each modality.

2. METHODS

The following mixed-effect linear decision model was used for the split-plot on reader design to generate raw data from different magnitudes for the variance components:

$$Y_{ijkt} = \mu_t + \tau_{it} + R(\tau)_{ijt} + C_{kt} + (\tau C)_{ikt} + \varepsilon_{(ijkt)} \quad ,$$

in which $\mu_t = 0$ if truth value (t) is negative or $\mu_t = a/b$ if truth value (t) is positive, where a and b are the population "location" and "scale" parameters, respectively, of the mean binormal ROC curve, τ_{it} is the fixed effect of modality i for truth value t, $R(\tau)_{ijt}$ is the random effect of reader j nested within modality i for truth value t, C_{kt} is the random effect of case k for truth value t, $(\tau C)_{ikt}$ is the random modality by case interaction effect for modality i, case k for truth value t, and $\varepsilon_{(ijkt)}$ is the random error associated with one reading defined by modality i, reader j, case k for truth value t.

The population ROC areas, latent variable structures, case-sample sizes and normal/abnormal case-sample ratios studied by Roe and Metz[3] were adapted for these simulations. Two changes were instituted for the split-plot on readers design. First, the variance component for reader-case interaction in the factorial design was combined with the residual in the split-plot on readers design. Second, the reader component of variance and the treatment by reader component of variance of the factorial design were summed to produce the readers nested within treatments component of variance in the split-plot on readers design.

Table 1 presents the analysis of variance for the split-plot design on readers using unrestricted parameterization. The lower part of Table 1 gives rules for selecting error terms to test treatment effects.

Table 1: Split Plot on Readers Analysis of Variance: Unrestricted Parameterization

Source	df	Expected Mean Square
Treatments (T)	1	$rc\sigma^2_{\tau} + c\sigma^2_{R(\tau)} + r\sigma^2_{\tau C} + \sigma^2_{R(\tau)C}$
Readers (Treatments)(R(T))	2(r−1)	$c\sigma^2_{R(\tau)} + \sigma^2_{R(\tau)C}$
Cases (C)	c−1	$2r\sigma^2_{C} + r\sigma^2_{\tau C} + \sigma^2_{R(\tau)C}$
T × C	c−1	$r\sigma^2_{\tau C} + \sigma^2_{R(\tau)C}$
R(T) × C	2(r−1)(c−1)	$\sigma^2_{R(\tau)C}$

Rules for Selecting Error Term to Test for Treatment Effects:

(i) If $MS_{R(\tau)} / MS_{R(\tau)\times C} \le 1$, and $MS_{\tau C} / MS_{R(\tau)\times C} \le 1$, use $MS_{R(\tau)\times C}$;
(ii) If $MS_{R(\tau)} / MS_{R(\tau)\times C} \le 1$, and $MS_{\tau C} / MS_{R(\tau)\times C} > 1$, use $MS_{\tau C}$;
(iii) If $MS_{R(\tau)} / MS_{R(\tau)\times C} > 1$, and $MS_{\tau C} / MS_{R(\tau)\times C} \le 1$, use $MS_{R(\tau)}$;
(iv) If $MS_{R(\tau)} / MS_{R(\tau)\times C} > 1$, and $MS_{\tau C} / MS_{R(\tau)\times C} > 1$, use Satterthwaite error term.

The following mixed-effect linear decision model was used for the split-plot on cases design to generate raw data from different magnitudes for the variance components:

$$Y_{ijkt} = \mu_t + \tau_{it} + C(\tau)_{ikt} + R_{jt} + (\tau R)_{ijt} + \varepsilon_{(ijkt)} \quad ,$$

in which $\mu_t = 0$ if truth value (t) is negative or $\mu_t = a/b$ if truth value (t) is positive, where a and b are the population "location" and "scale" parameters, respectively, of the mean binormal ROC curve, τ_{it} is the fixed effect of modality i for truth value t, $C(\tau)_{ikt}$ is the random effect of case k nested within modality i for truth value t, R_{jt} is the random effect of reader j for truth value t, $(\tau R)_{ijt}$ is the random modality by reader interaction effect for modality i, reader j for truth value t, and $\varepsilon_{(ijkt)}$ is the random error associated with one reading defined by modality i, reader j, case k for truth value t.

Two changes were instituted for the split-plot on cases design. First, the variance components for cases and treatment by cases of the factorial design were combined to produce the component of variance for cases nested within treatments. Second, the reader-by-case interaction and the residual of the factorial design were summed to produce the residual component of variance for the split-plot on cases design.

Table 2 presents the analysis of variance for the split-plot design on cases using unrestricted parameterization. The lower part of Table 2 gives rules for selecting error terms to test treatment effects.

Table 2: Split Plot on Cases Analysis of Variance: Unrestricted Parameterization

Source	df	Expected Mean Square
Treatments (T)	1	$cr\sigma^2_{\tau} + r\sigma^2_{C(\tau)} + c\sigma^2_{\tau R} + \sigma^2_{C(\tau)R}$
Case (Treatments)(C(T))	2(c−1)	$r\sigma^2_{C(\tau)} + \sigma^2_{C(\tau)R}$
Reader (R)	r−1	$2c\sigma^2_{R} + c\sigma^2_{\tau R} + \sigma^2_{C(\tau)R}$
T × R	r−1	$c\sigma^2_{\tau R} + \sigma^2_{C(\tau)R}$
C(T) × R	2(r−1)(c−1)	$\sigma^2_{C(\tau)R}$

Rules for Selecting Error Term to Test for Treatment Effects:

(i) If $MS_{C(\tau)} / MS_{C(\tau)\times R} \le 1$, and $MS_{\tau R} / MS_{C(\tau)\times R} \le 1$, use $MS_{C(\tau)\times R}$;
(ii) If $MS_{C(\tau)} / MS_{C(\tau)\times R} \le 1$, and $MS_{\tau R} / MS_{C(\tau)\times R} > 1$, use $MS_{\tau R}$;
(iii) If $MS_{C(\tau)} / MS_{C(\tau)\times R} > 1$, and $MS_{\tau R} / MS_{C(\tau)\times R} \le 1$, use $MS_{C(\tau)}$;
(iv) If $MS_{C(\tau)} / MS_{C(\tau)\times R} > 1$, and $MS_{\tau R} / MS_{C(\tau)\times R} > 1$, use Satterthwaite error term.

For each type of split plot design, a series of null-case Monte Carlo simulations were conducted with two modalities. For the split-plot design on readers, 3, 5, and 10 different hypothetical readers were nested within each modality, and 10+/90−, 25+/25−, 50+/50−, and 100+/100−cases were crossed with modalities. For the split-plot design on cases, 10+/90−, 25+/25−, 50+/50−, and 100+/100−cases were nested within each modality, and 3, 5, and 10 hypothetical readers were crossed with modalities.

Our series of null-case computer simulations were conducted to examine the relation between the nominal type I error rate and the empirical type I error rate with five-category discrete rating data as recommended by Swets and Pickett.[4] Two thousand samples were generated for each condition. In the computer simulation, a continuous decision variable was generated by assuming a linear mixed model for the decision variable comparable to the linear mixed model for the jackknife pseudovalues. Roe and Metz[3] and Dorfman, Berbaum, Lenth et al.[5] used the equal-variance binormal model ($b = 1$). Therefore, in all of the Monte Carlo simulations that follow, we also used the equal-variance binormal model to

facilitate comparison between our results and their results.

The null hypothesis was true in all simulations, and the population ROC curves were the same for both modalities. Three binormal population ROC curves with A_z values of .702, .855, and .961, corresponding to $\mu_{pos} = 0.75$, 1.5, and 2.5, were included in the study. The magnitudes of the decision-variable variance components were chosen to be the same for actually-positive and actually-negative cases. The decision thresholds were set at 0.25, 0.75, 1.25, and 1.75 corresponding to false-positive fractions of 0.40, 0.23, 0.11, and 0.04. The decision thresholds serve mathematically as the values of the category boundaries for the continuous random decision variable. The decision thresholds were chosen to mirror the moderate conservatism often observed in radiologic image interpretation.

3. RESULTS

For equal allocation ratios, small (A_z=0.702) and moderate ROC area (A_z=0.855), empirical Type I error rate closely matched nominal alpha level; however, for very large ROC area (A_z=0.961), empirical Type I error rate was somewhat smaller than nominal alpha level. For the most part, these findings are consistent with those reported by Roe and Metz,[3] and by Dorfman, Berbaum, Lenth, et al.[5]

Figures 1, 2, and 3 present the results for areas of .702, .855, and .961, respectively, with the upper panel representing the split-plot design on readers and the lower panel representing the split-plot design on cases. To facilitate comparison with Monte Carlo studies based on the factorial design, we used the same labels for the variance structures as those employed by Roe and Metz,[3] and Dorfman, Berbaum, Lenth et al.[5] In these figures, the 95% probability band about the nominal alpha level is shown as the band between the two horizontal lines at 0.040 and 0.060 for $\alpha = 0.05$. The boundaries we selected are standard critical values for 95% probability bands for a binomially distributed random variable derived from the normal approximation to the binomial distribution when the sample size n is large and the binomial probability p is known.[6] The boundaries define a 5% rejection region for the null hypothesis that alpha is a specified value, and that empirical Type I error rate follows a binomial distribution with parameters n = 2000 and $p = 0.05$.

In this paper, we focus on equal allocation ratios. The split-plot design on readers appears to perform somewhat better than the split-plot design on cases. The top panel of Figures 1 and 2 ($A_z = 0.702$ and 0.855, split-plot on readers) shows that the Monte Carlo data points fell, for the most part, within the 95% probability band when the number of cases was at least a hundred (50+/50−), or the number of readers per modality was at least ten. With fewer than a hundred cases and fewer than ten readers per modality, the statistical test was slightly conservative. The bottom panel of Figures 1 and 2 ($A_z = 0.702$ and 0.855, split-plot on cases) shows that the Monte Carlo data points fell, for the most part, within the 95% probability band when there are ten readers. With fewer than ten readers, the statistical test was slightly conservative.

Figure 3 ($A_z = 0.961$) shows that the Monte Carlo data points almost always fell outside the 95% probability band. For both split-plot designs, 200 cases was sufficient to keep the points either within the band or close to it. With 100 cases, the test was somewhat conservative; with 50 cases, the test was quite conservative.

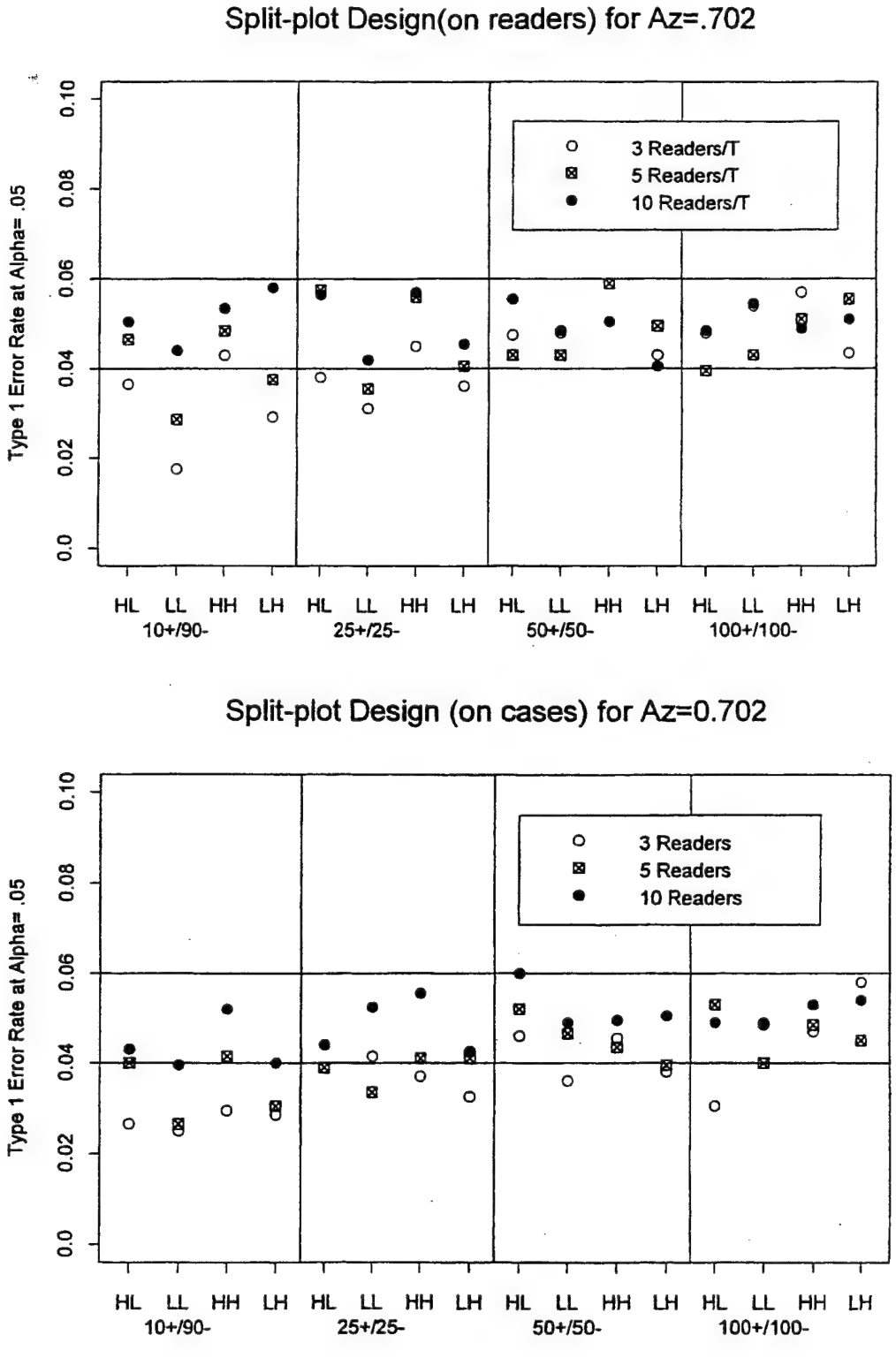


Figure 1: Split-plot results on readers (top) and cases (bottom) for A_z=.702, and nominal α=.05.

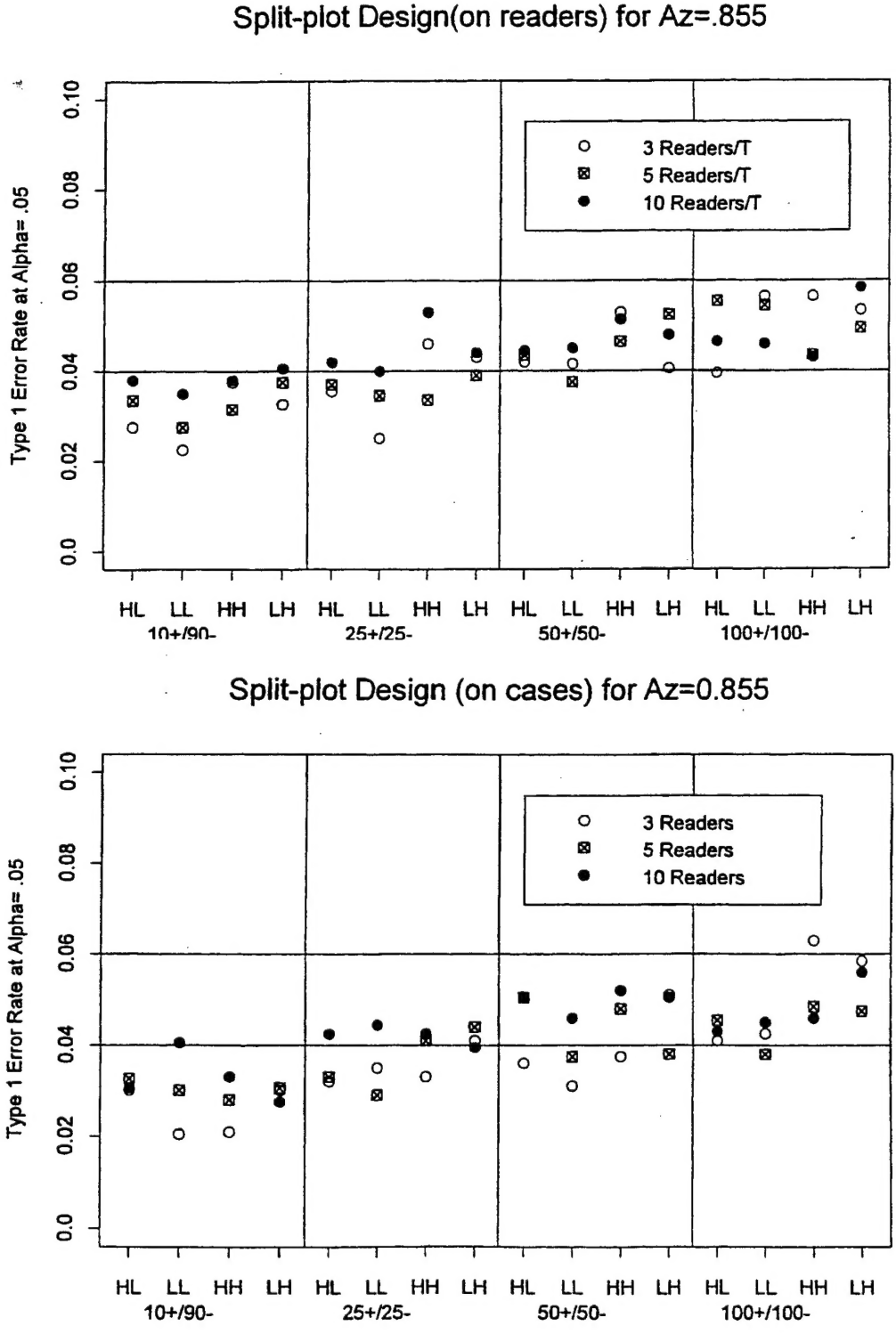


Figure 2: Split-plot results on readers (top) and cases (bottom) for A_z=.855, and nominal α=.05.

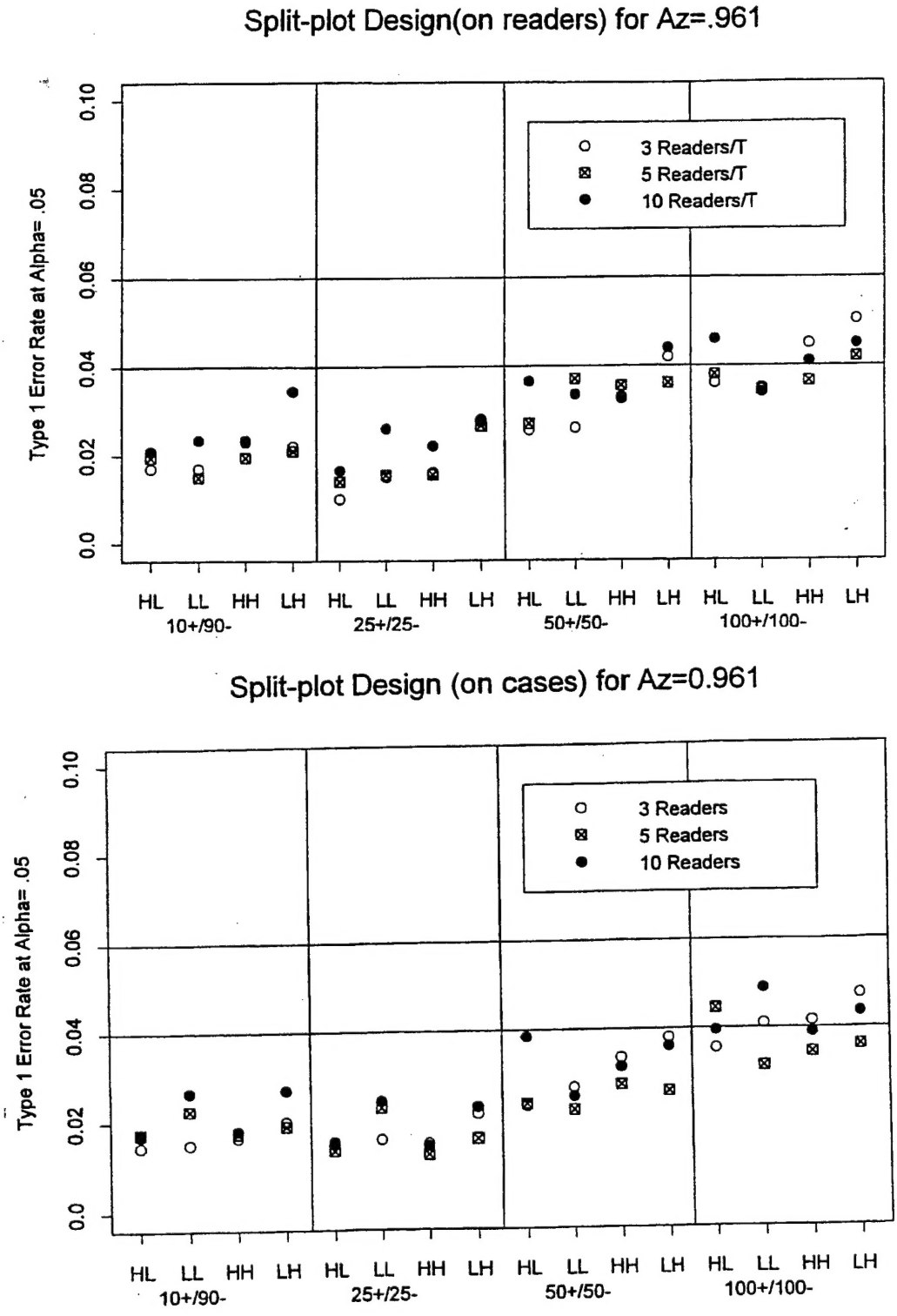


Figure 3: Split-plot results on readers (top) and cases (bottom) for A_z=.961, and nominal α=.05.

4. DISCUSSION

Our Monte Carlo simulations show that the DBM multireader method can be validly extended to the reader-split and case-split plot designs. The results suggest that the DBM method provides trustworthy alpha levels with discrete ratings when ROC area is not too large, and case and reader sample sizes are not too small. In other situations, the test tends to be conservative. A statistical test is conservative if the empirical Type I error rate level is smaller than the nominal alpha level. If one rejects the null hypothesis at a specified nominal alpha level, a conservative statistical test is preferred to a liberal test because it has a lower Type I error rate. The Monte Carlo data showed that the statistical test generally becomes more conservative with large area and decreasing case sample size.

In our previous Monte Carlo validation of the DBM multireader method in a fully crossed factorial design as well as in these reader and case split-plot studies, we used discrete rating scales. Pseudo-continuous rating scales have been recommended over discrete scales for routine use in ROC studies in diagnostic radiology primarily because it was concluded that pseudo-continuous rating scales were less likely to yield binormal degenerate data sets than discrete rating scales.[7,8] Binormal degenerate data sets are no longer an issue in ROC analysis.[9-11] Moreover, the empirical evidence suggests that discrete and pseudo-continuous scales can often be used interchangeably in image evaluation studies when the investigator is interested in ROC area because they produce virtually the same results.[7] If, however, the experimenter is interested in the operating points as well ROC area, then discrete rating scales should be used.[5] Sensitivity and specificity are determined by the location of the operating points on the ROC curve as well as by the discriminability of the underlying distributions of normal and abnormal cases on the latent decision dimension. The operating points are determined by decision thresholds, and in clinical trials, these decision thresholds are, in fact, action thresholds. For instance, in the American College of Radiology Breast Imaging Reporting and Data System (BI-RADS), "probably benign finding" translates into the course of action "short interval followup suggested," "suspicious abnormality" translates into the course of action "biopsy should be considered," and "highly suggestive of malignancy" translates into "appropriate action should be taken".[12] Some diagnostic imaging systems may lead to more conservative or liberal actions than others. For example, a concern with algorithms for computer-aided diagnosis (CAD) is that they seem to give too many false positives.[13] If the radiologist's decision thresholds are changed by the rate of CAD false positives, changes in sensitivity and specificity would occur, but would not be observed with LABROC-type algorithms that sort pseudo-continuous rating data into discrete uniformly distributed categories.[14,15] Because the category boundaries defined by such algorithms do not correspond to the decision thresholds used by the observers both within and between modalities, conclusions about radiologist performance might be drawn based solely on the ROC area, while inappropriately ignoring the differences in true and false positive fractions associated with the observers' operating points.[5]

5. CONCLUSIONS

Our Monte Carlo simulations show that the DBM multireader method can be validly extended to reader-split and case-split plot designs.

6. ACKNOWLEDGMENTS

Supported in part by National Institutes of Health grants R01 CA 62362 (D.D.D., K.S.B., R.V.L., Y-F.C.), and R01 CA 42453 (D.D.D and K.S.B.) and in part by US Army Medical Research and Materiel Command grant DAMD17-96-1-6254 (D.D.D., K.S.B., R.V.L.).

7. REFERENCES

1. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. Invest Radiol 1992; 27:723-731.
2. Obuchowski NA, Zepp RC. Simple steps for improving multiple-reader studies in radiology. AJR 1996; 166:517-521.
3. Roe CA, Metz CE. The Dorfman-Berbaum-Metz method for statistical analysis of multi-reader, multi-modality ROC data: Validation by computer simulation. Acad Radiol 1997; 4:298-303.
4. Swets JA, Pickett RM. Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. New York, NY: Academic Press, 1982.
5. DD Dorfman, KS Berbaum, RV Lenth, Y-F Chen, BA Donaghy. Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: Factorial experimental design. Academic Radiology 1998;5:591-602.
6. Snedecor GW, Cochran WG. Statistical Methods. (Eighth Edition). Ames, IA: Iowa State University, 1989.
7. Rockette HE, Gur D, Metz CE. The use of continuous and discrete confidence judgments in receiver operating characteristic studies of diagnostic imaging techniques. Invest Radiol 1992; 27:169-172.
8. King JL, Britton CA, Gur D, Rockette HE, Davis PL. On the validity of the continuous and discrete confidence rating scales in receiver operating characteristic studies. Invest Radiol 1993; 28:962-963.
9. Dorfman DD, Berbaum KS. Degeneracy and discrete ROC rating data. Acad Radiol 1995; 2:907-915.
10. Dorfman DD, Berbaum KS, Metz CE, Lenth RV, Hanley JA, Abu-Dagga H. Proper receiver operating characteristic analysis: The bigamma model. Acad Radiol 1997; 4:138-149.
11. Pan X, Metz CE. The "proper" binormal model: Parametric receiver operating characteristic curve estimation with degenerate data. Acad Radiol 1997; 4:380-389.
12. Breast imaging - reporting and data system (BI-RADS). Reston, VA.: American College of Radiology, 1993.
13. Yoshida H, Doi K, Nishikawa RM, Giger ML, Schmidt RA. An improved computer-assisted diagnostic scheme using wavelet transform for detecting clustered microcalcifications in digital mammograms. Acad Radiol 1996; 3:621-7.
14. Metz CE, Shen JH, Herman BA. New methods for estimating a binormal ROC curve from continuously-distributed test results. Presented at the 1990 Joint Statistical Meetings of the American Statistical Association and the Biometric Society, Anaheim, CA, August, 1990.
15. Vittitoe NF, Baker JA, Floyd CE Jr. Fractal texture analysis in computer-aided diagnosis of solitary pulmonary nodules. Acad Radiol 1997;4:96-101.